

Saif Hameed
Zeeshan Fatima *Editors*

Integrated Omics Approaches to Infectious Diseases

 Springer

Integrated Omics Approaches to Infectious Diseases

Saif Hameed • Zeeshan Fatima
Editors

Integrated Omics Approaches to Infectious Diseases

 Springer

Editors

Saif Hameed
Amity Institute of Biotechnology
Amity University Haryana
Gurugram, Haryana, India

Zeeshan Fatima
Amity Institute of Biotechnology
Amity University Haryana
Gurugram, Haryana, India

ISBN 978-981-16-0690-8

ISBN 978-981-16-0691-5 (eBook)

<https://doi.org/10.1007/978-981-16-0691-5>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

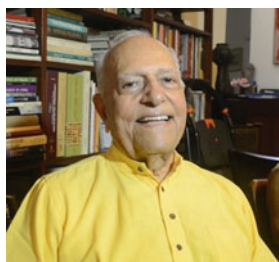
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

*Dedicated to our
“Loving Daughters”*



Foreword



G. P. Talwar
Docteurs Sciences, DSc (hc), FAMS, FASc, FNASc,
FNA, FRCOG, FWAAS
Talwar Research Foundation, New Delhi, India

A new era has dawned! Here is a book compiled by two mid-career scientists, Saif Hameed and Zeeshan Fatima, on the unusual theme of Integrated Omics. The field explored is authored by active working research scientists. The chapters provide valuable information and progress at molecular level in several domains relating to pathogenesis and malfunction caused by microorganisms.

Genomics, proteomics, peptidomics, metabolomics, lipidomics, translational omics, and pathogen-omics are among the terms used and elaborated in their chapters by the contributing authors; hence, the emergence of the “Novel” title of the book.

Let me hope that this book finds “readers” beyond the mid-career, active, and busy scientists, who no doubt would be the noteworthy component of science in India and elsewhere in the world in the coming years.

G. P. Talwar

Foreword



V. M. Katoch, MD, FNASc, FASc, FAMS, FNA
Rajasthan University of Health Sciences (RUHS),
Jaipur, India
JIPMER, Puducherry, India
AIIMS, Madurai, India
Department of Health Research, Government of India,
Jaipur, Rajasthan, India
Indian Council of Medical Research, Jaipur, Rajasthan,
India

I am extremely delighted to write this foreword for the book entitled “Integrated Omics Approaches to Infectious Diseases” being compiled and edited by Dr. Saif Hameed and Dr. Zeeshan Fatima.

This book attempts to present a wide view of the current status of integrative omics approaches to collate recent developments in this field into a state-of-the-art framework to understand human infection biology. This book has covered major aspects related to the use of OMICs approaches to understand the structure–function relationships between infectious agents and their hosts with a major focus on pathobiology of different infectious diseases. Knowledge generated by the application of different technologies and approaches described will be relevant to many stakeholders in progressing towards the development of better diagnostics, newer drugs and vaccines to better treat, control, and eliminate these diseases. Both the editors and contributing authors are very distinguished experienced colleagues from India and other countries with original contributions to their credit. This is reflected by their insight in identification of needs, description of technologies, and highlighting their application(s).

This book has quite informative sections on infection genomics: genomics applications for human pathogens; infection transcriptomics: transcriptomics applications for human pathogens; infection proteomics: proteomics applications for human pathogens; infection metabolomics: metabolomics applications for

human pathogens, and translation omics: application of omics in translational research. 28 chapters covered under these 5 broad sections deal with almost all relevant technologies important for studying genetics, epigenetics, structural and functional genomics as well as proteomic analysis of important pathogens as well as their interactions with host. Though only a limited number of organisms have been described, these should be considered as examples. However, in my opinion the book will help all biologists interested in human medicine, veterinary medicine, and even plant diseases. Application of these generic technologies and approaches will help the readers to think of newer strategies/targets for developing easy to implement cost-effective tools for diagnosis, effective therapeutics, and vaccines for various infectious diseases.

I compliment the authors as well as editors for this excellent effort. I am optimistic that this research-based reference book will be beneficial for biomedical scientists/ teachers, researchers, and health care industries involved in various aspects of infectious diseases. I am also sure that this book will be liked by students belonging to diverse disciplines with interest on different aspects of contemporary and emerging as well as reemerging infectious diseases.



V. M. Katoch

Preface

Omic refers to “*the biology propelled by technology*”

Infectious diseases have posed resilient challenges to humans since ages, and it is still more complicated by the emergence of antimicrobial resistance. Now, COVID-19 pandemic made the world realize the ruthless devastating effects of human pathogens on not only health but all facets of human life. Scientific community is constantly encountering the challenges put forth by infectious pathogens and burdened finding out the novel means to transgress their impact on society. To gain a leap over these pathogens and to better understand their pathogenesis, development of novel technologies is the need of the hour. “Omics” approaches are one of the advanced technology-driven tools that have presented a potential alternative and aided understanding at the molecular level.

Omic studies have revolutionized the various aspects of molecular biology by facilitating high-throughput comprehensive appraisal of molecules in a biological system. It comprises genomics—profiling of DNA, transcriptomics—profiling of RNA, proteomics—profiling of proteins, and metabolomics—profiling of metabolites. The holistic understanding of biological molecules requires the understanding and integration of the omics technology in the form of “multi-omics” and “meta-omics” by extrapolating omics circumferences to the epigenome and microbiome. From genomics to transcriptomics and proteomics now we are heading toward metabolomics. The journey has started since the identification of single gene product, but the power of omics has dissected the avenues for study of non-targeted genes and gene products with their dynamics and influence on their surrounding environment.

Integrated Omics Approaches to Infectious Diseases is a unique reference book catering to “omics”-based technological advancements and current research in the field of infectious diseases. This book covers omics studies on major microbes, namely bacteria, fungi, protozoa, and virus causing infections in humans, on a single platform and summarizes microbial pathogenesis, drug resistance, diagnostics, vaccines, and novel therapeutic aspects of various microbial life-threatening diseases. The book is divided into five clearly structured sections comprising

genomics, transcriptomics, proteomics, metabolomics, and translational omics which are the major branches of omics technology further extrapolated to “epi-,” “multi-,” and “meta-” omics. The infection genomics sections deal with applications of genomics, epigenomics, next-generation sequencing, and bioinformatics in virulence, host–pathogen interactions, drug resistance, and diagnostics of bacterial and viral pathogens. The infection transcriptomics section deals with the application of transcriptomics approaches such as microarray, noncoding RNA profiling, RNA sequencing, and miRNA in bacterial, fungal, and viral pathogens. The infection proteomics section deals with proteomics applications in understanding the molecular mechanism of antibacterial drugs and identification of drug targets. Additionally, one chapter on proteomics application to study post-translational mechanisms in parasite is also included with special emphasis on phosphorylation and methylation events. The infection metabolomics section describes the recent addition of another omics branch, i.e., metabolomics to study disease biomarkers, host–pathogen interaction, and lipid droplets in protozoan and fungal pathogens. The final section of translational omics describes the applications of omics technology in translational research such as nanotechnology. The last chapter written by the editors themselves describes various challenges and future prospects of omics technology to study human pathogens.

Furthermore, each chapter layout has an introduction to the current state followed by application of that technology and finishing with its future prospects. The book is contributed by the eminent veterans and selected experts of international repute working in diverse areas of omics technology. It is their cumulative efforts that discuss cutting-edge omics research in their respective areas of infectious diseases dealing with a comprehensive description of recent developments and depicting the future leads. The research described in this book may further be extended to other pathogens and chronic and metabolic disorders. The book chapters are presented in a way accessible to the teachers, UG/PG students, research scholars, and healthcare workers including pathologists and clinicians worldwide.

We are grateful for the blessings and constant motivation from Dr. Ashok K. Chauhan, President RBEF, and Dr. Aseem K. Chauhan, Additional President RBEF and Chancellor, Amity University Haryana (AUH) and Amity University Rajasthan (AUR). Sincere thanks to Dr. P.B. Sharma, VC, AUH, Maj. Gen. B.S. Suhag, DVC, AUH, Dr. Padmakali Banerjee, PVC, AUH, and Dr. Rajendra Prasad, Dean Research, AUH, for their overall support to enhance our academic rigor. We are grateful to our esteemed contributors for their worthy and timely contributions during the tough times of global pandemic without which this compilation would not have become a ready reference for the researchers in this field. We take pride in thanking two living legends Padma Bhushan Prof. G. P. Talwar and former Director-General, ICMR, and Secretary, Department of Health, Government of India, Dr. V. M. Katoch for endorsing this book by giving their valuable forewords. Patience and support from Dr. Bhavik Sawhney, Springer Nature, during

the book preparation is deeply acknowledged. Last but not the least, we dedicate this piece of work to our loving twin daughters (Aima Zainab and Aiza Maryam) whose share of precious childhood time was used for our research endeavors.

Gurugram, Haryana, India

Saif Hameed
Zeeshan Fatima

Contents

Part I Infection Genomics: Genomics Applications for Human Pathogens

1	Geno-informatics for Prediction of Virulence and Drug Resistance in Bacterial Pathogens	3
	Umay Kulsum, Praveen Kumar Singh, S. Rashmi Mudliar, and Sarman Singh	
2	Next Generation Sequencing: Opportunities and Challenges in Tuberculosis Research	19
	Faraz Ahmad, Anwar Alam, Indu Kumari, Sugandha Singh, Anshu Rani, Aquib Ehtram, Soumya Suhasini, Jasmine Samal, and Nasreen Z. Ehtesham	
3	Deciphering the Role of Epigenetic Reprogramming in Host-Pathogen Interactions	41
	Amandeep Kaur Kang, Andrew M. Lynn, and Uma Dhawan	
4	Genomic Evidence Provides the Understanding of SARS-CoV-2 Composition, Divergence, and Diagnosis	63
	Manish Tiwari, Gurparsad Singh Suri, Gurleen Kaur, Baljinder Singh, Sahil Mehta, and Divya Mishra	
5	Importance of Next-Generation Sequencing in Viral Diagnostics	81
	Ashish Kumar Vyas and Sudheer Gupta	
6	Bioinformatic Application in COVID-19	87
	Gurjot Kaur, Soham Mukherjee, and Shreya Jaiswal	

Part II Infection Transcriptomics: Transcriptomics Applications for Human Pathogens

- 7 Emerging Transcriptomic Approaches to Decipher Mycobacterial Complexities** 107
 Jasmine Samal, Nilofer Naqvi, Yashika Ahuja, Neha Quadir, P. Manjunath, Faraz Ahmad, Mohd. Shariq, Anwar Alam, Avantika Maurya, and Nasreen Z. Ehtesham
- 8 Microarrays: A Road Map to Uncover Host Pathogen Interactions** 125
 Heerak Chugh, Gagan Dhawan, Ramesh Chandra, and Uma Dhawan
- 9 Transcriptional Approach in the Identification of Drug Targets in *Candida* spp.** 139
 Mahnoor Patel, M. Amin-ul Mannan, and Banhishikha Datta
- 10 Noncoding RNA Profiling: Potential Application in Infectious Diseases** 157
 Shiffali Khurana, Uma Dhawan, and Vibha Taneja
- 11 RNA-Seq Analysis Strategies to Understand Viral Pathogenesis** . . . 185
 Anvitha Nair, Arpana Vibhuti, V. Samuel Raj, and Ramendra Pati Pandey
- 12 Various Transcriptomic Approaches and Their Applications to Study Small Noncoding RNAs in Dengue and Other Viruses** . . . 195
 Deeksha Madhry, Kush Kumar Pandey, Shivani Malvankar, Shubham Kumar, Anjali Singh, Ravi Kumar S. Yelegara, Rupesh K. Srivastava, and Bhupendra Verma
- 13 Transcriptomic Approaches in Understanding SARS-CoV-2 Infection** 221
 Sona Charles and Jeyakumar Natarajan
- 14 miRNA Target Prediction: Overview and Applications** 241
 Fazlur Rahman, Sajjadul Kadir Akand, Muniba Faiza, Shams Tabrez, and Abdur Rub

Part III Infection Proteomics: Proteomics Applications for Human Pathogens

- 15 A Glimpse into Peptidomic Approach** 257
 V. S. Gowri and V. Sabareesh

16	Molecular Mechanism of Action of Antimicrobial Agents Against Clinically Important Human Pathogens: A Proteomics Approach . . .	287
	Anthonyimuthu Selvaraj, Alaguvel Valliammai, and Shunmugiah Karutha Pandian	
17	Exploration of the Mycobacterial Proteome in the Pathogenesis of TB: A Perspective	303
	Mohd. Shariq, Sheeba Zarin, Nilisha Rastogi, Indu Kumari, Farha Naz, Tarina Sharma, Neha Sharma, and Nasreen Z. Ehtesham	
18	Pathogenesis of <i>Staphylococcus aureus</i> and Proteomic Strategies for the Identification of Drug Targets	325
	Alaguvel Valliammai, Anthonyimuthu Selvaraj, and Shunmugiah Karutha Pandian	
19	Proteomics in the Study of Host-Pathogen Interactions	341
	Preethi Sudhakara, S. Kumaran, and Wilson Aruni	
20	Significance of Post-translational Modifications in Apicomplexan Parasites	359
	Priya Gupta, Rashmita Bishi, Sumbul Khan, Avi Rana, Nirpendra Singh, and Inderjeet Kaur	
Part IV Infection Metabolomics: Metabolomics Applications for Human Pathogens		
21	An Introduction to Computational Pipelines for Analyzing Untargeted Metabolomics Data for Leishmaniasis	375
	Anita Verma, Arunangshu Das, and Chinmay K. Mukhopadhyay	
22	Metabolomics: A Promising Tool to Study Disease Biomarkers and Host-Pathogen Interactions	403
	Megha, Preeti, and Tulika Prasad	
23	Lipidomics to Study the Role of Lipid Droplets in Host-Pathogen Interactions	425
	Anwesha Bhattacharyya and Vineet Choudhary	
24	Lipid Structure, Function, and Lipidomic Applications	441
	Khusboo Arya, Sana Akhtar Usmani, Nitin Bhardwaj, Sudhir Mehrotra, and Ashutosh Singh	
Part V Translational Omics: Omics Applications in Translational Research		
25	Nanoparticles as Therapeutic Nanocargos Affecting Epigenome of Microbial Biofilms	461
	Indu Singh, Pradeep Kumar, and Gagan Dhawan	

26	Malaria in the Era of Omics: Challenges and Way Forward	483
	Manish Tripathi, Amit Khatri, Vaishali Lakra, Jaanvi Kaushik, and Sumit Rathore	
27	Omics Approaches for Infectious Diseases	507
	Amrendra Nath Pathak, Lalit Kumar Singh, and Esha Dwivedi	
28	Pathogen-Omics: Challenges and Prospects in Research and Clinical Settings	521
	Dyuti Purkait, Saif Hameed, and Zeeshan Fatima	

About the Editors

Saif Hameed is currently an Associate Professor at Amity Institute of Biotechnology, Amity University, Haryana. Dr. Hameed did his Bachelor's from the University of Delhi and Master's from Jamia Hamdard and completed his doctoral studies in Life Sciences from Jawaharlal Nehru University. He also worked as a visiting scholar in Institute für Mikrobiologie, Heinrich-Heine-Universität, Düsseldorf, Germany, in 2008. Dr. Hameed has received the Young Scientist award from the Science and Engineering Research Board, Department of Science and Technology, New Delhi. Dr. Hameed is actively engaged in research in infectious diseases particularly multidrug resistance (MDR) in pathogenic fungi. He has around 55 peer-reviewed papers to his credit in high repute journals, 5 books, 22 book chapters, and 25 popular articles, with 2 of them as 'cover story', and guided 4 doctoral students. He is a life member of the Association of Microbiologists of India (AMI) and the International Society for Infectious Diseases (ISID).

Zeeshan Fatima is currently working as Associate Professor at Amity Institute of Biotechnology, Amity University, Haryana. Dr. Fatima did her Bachelor's and Master's from Banaras Hindu University and earned her doctoral degree in Biochemistry from Aligarh Muslim University. She has held research positions under nationally and internationally funded research projects, including her Research Associateship at BHU and JNU and postdoctoral training from the University of Cincinnati, Ohio, USA. Dr. Fatima has received two Young Scientist awards under Fast Track and Women Scientist Schemes, respectively, from the Science and Engineering Research Board, Department of Science and Technology, New Delhi. She is actively engaged in infectious diseases research, particularly on multidrug resistance in human pathogens, *Mycobacterium tuberculosis* and *Candida albicans*. She has 5 books and more than 50 peer-reviewed papers to her credit in journals of high repute and supervised 4 doctoral thesis. She is a life member of the Association of Microbiologists of India (AMI), Society of Biological Chemists (SBC), India and the International Society for Infectious Diseases (ISID).

Part I

**Infection Genomics: Genomics Applications for
Human Pathogens**



Geno-informatics for Prediction of Virulence and Drug Resistance in Bacterial Pathogens

1

Umay Kulsum, Praveen Kumar Singh, S. Rashmi Mudliar, and Sarman Singh

Abstract

In recent years, there has been a rapid surge in the number of resistant strains of bacterial pathogens, mainly due to misuse of antibiotics in our day-to-day environment for treating common ailments. Hence new drug discovery is a priority. The discoveries of new antimicrobials increasingly rely on genotypic data resulting from whole genome sequencing. In recent years, there has been advancement in the whole genome sequence facilities and a number of completed genomes, but the wealth of information is not being fully utilized. Therefore, there is a need of microbial informatics algorithms to exploit this genomic data and provide an opportunity for the development of newer remedies. The combined advent of both genomics and bioinformatics can help in the identification, screening, and refinement of drug targets and predict drug resistance leading towards fast and efficient antimicrobial therapeutics. Therefore, this chapter is focused on the *in silico* approaches which can facilitate understanding, identifying, and controlling the virulence and bacterial antibiotic resistance.

Keywords

Drug resistance · Bioinformatics · Genomics · Virulence factors

U. Kulsum · S. R. Mudliar · S. Singh (✉)
Molecular Medicine Laboratory, Department of Microbiology, All India Institute of Medical Sciences, Bhopal, Madhya Pradesh, India

P. K. Singh
Department of Surgery, Millers School of Medicine, University of Miami, Coral Gables, FL, USA

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_1

1.1 Introduction

Antibiotic resistance is one of the major threats to global health estimating 70,000 deaths annually and is predicted to reach a toll of ten million by 2050 if not intervened [1]. Despite the alarming situation, antibiotics are still important not only to treat bacterial infections but also required in combination with most of modern medicines. The rapid increase in antibiotic resistance in recent years has left us with fewer antibiotics, yet the research community is unable to take complete advantage of genome-scale tools. Phenotypic methods like culture-based antimicrobial susceptibility testing (AST) are still the primary method, which dominates the epidemiology and effects of antibiotics. *In vitro* measurements are used to generate antibiotic susceptibility data [2]. These comprise mainly of measuring minimum inhibitory concentrations (MICs) of antibiotics using several-fold serial dilutions or the diameter of inhibition zones around disks containing a standard amount of antibiotics. Based on these techniques appropriate anti-infective therapy is selected. Although these methods are widely used, they have some well-known disadvantages such as shifting of interpretive standards, lack of valid methods for most organisms, and limitations of the number of agents that can be tested [3]. Many studies have been using PCR and microarrays for the detection of required specific genes [4]. These techniques can be used for identifying virulent determinants as well [5]. Identifying specific genes for a large number of isolates is not only laborious and costly but also less informative as it can provide information related to that particular gene only. Whole genome sequencing (WGS) has emerged as an alternative. Genotypic methods rely on the identification of antibiotic resistance genes and mutations and can be considered more suitable for detecting resistance. WGS technology has become a tool ideal for surveillance as it provides an affordable evaluation of the entire genome of a bacterium within few days at comparatively low costs [6]. The full complement of resistant determinants can be achieved by characterizing a microbe using WGS [7]. It also includes resistant determinants of compounds not phenotypically tested. In addition, WGS along with bioinformatics analysis also assists in identifying and deciphering the virulence potential in clinical isolates which can lead to the development of novel drugs, vaccines, and molecular diagnostic tools for improving therapeutic management.

Bioinformatics is an interdisciplinary field of science employing a range of computational techniques along with mathematics and statistics for biological data interpretation involving sequence, structural alignment, and analysis of genomics, transcriptomics, and proteomics dataset [8]. It has created an opportunity from high-throughput data and computational techniques to get insights of bacterial drug resistance. The available computational resources and the open-source software have made the integration of multi-omics data to enhance knowledge of therapeutic doses. Bioinformatics studies have provided an important system of modeling which involves combining the molecular understanding of bacterial systems. Moreover, databases and literature for producing molecular profiles and for gathering information related to the epidemiological study of pathogens have essentially expanded. Hence, the utilization of bioinformatics tools and methods in controlling microbial

resistance, identification of pathogens, and recognizing markers for early diagnosis and treatment is essential [9]. This chapter provides an overview of bacterial virulence factors and antimicrobial resistance mechanism with the main focus on the bioinformatics approaches for identifying virulence and antimicrobial resistance which can help in screening and refinement of drug target candidates, thus facilitating newer therapeutics and disease management.

1.2 Deciphering the Virulence Repertoire of a Bacterial Strain

Virulence is referred to as the ability of a pathogen to cause diseases depending on various factors termed as “virulence factors.” Bacterial virulence factors are a certain set of genes which help the bacteria to survive under selection pressures and adapt to a new environment. These factors include secretory proteins (toxins, exoenzymes, type I to VI secretion systems), biofilm-forming proteins, siderophores, as well as polysaccharides [10]. Lipopolysaccharide (LPS), a component of the outer cell membrane, is also a virulence factor in Gram-negative bacteria. The bacterial pathogens use these factors as an equipment to establish infection by invading cells and confronting host defense mechanisms. Bacterial virulence factors are located on the pathogenicity islands and are most often acquired by horizontal gene transfer (HGT) [11, 12]. Earlier virulence factors were identified using biochemical approaches such as purifying the virulence factor genes and studying their effect *in vitro* or *in vivo*. Another approach used was the screening of a panel of genes playing a role in pathogenesis by mutagenesis or molecular cloning and expression in non-pathogenic strains [13, 14]. Since the last two decades, the discovery rate of virulence factors has increased dramatically with the help of genomics along with transcriptomics and proteomics. The genomic methods for identification of virulence factors involve homology search with known virulence genes, comparing strains with various levels of virulence, or analyzing the horizontally acquired genes [15, 16].

1.3 Antibiotic Resistance in Bacterial Pathogens

Antibiotic resistance in bacteria depends on the dose of exposure to antibiotics [17]. For instance, even the most resistant bacteria can be killed or inhibited by a high concentration of antibiotics, but a patient might not be able to tolerate the high concentration. The concentration of antibiotics depends on the bacteria’s susceptibility profile. There are two concepts for antibiotic resistance: first is acquired resistance by the bacteria, and the other is *in vitro* activity of antibiotics against a bacterial population. Acquired resistance is a result of the bacteria trying to adapt to certain environmental conditions for its survival. The bacteria usually acquire resistance either due to mutations in chromosomal genes or due to genetic determinants of resistance acquired from external intrinsically resistant organisms present in the environment (generally by HGT). Bacterial antibiotic resistance also

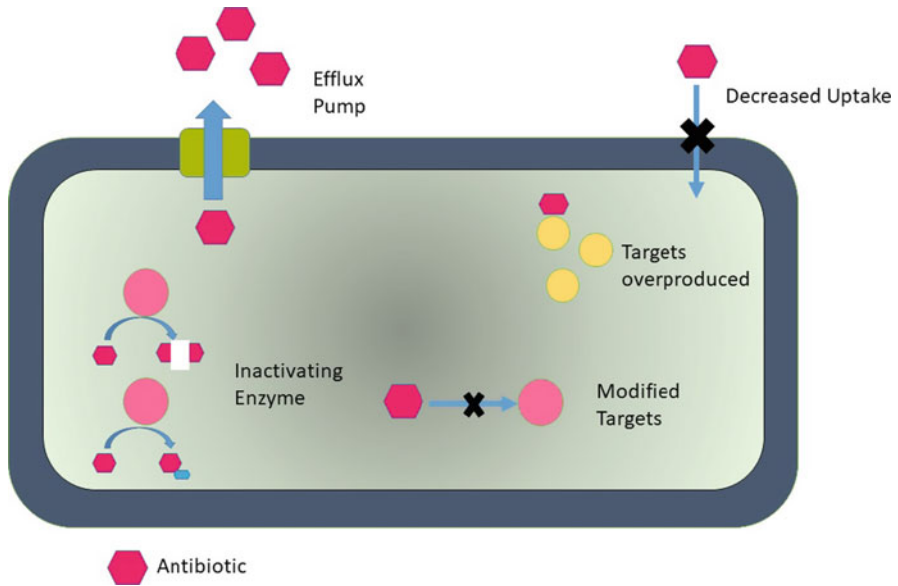


Fig. 1.1 Common mechanism of bacterial resistance against antibiotics

depends on the adaptive mechanism such as structural modification of drug molecule [18]. Another protective mechanism exhibited by bacteria is expelling the antibiotic out of the bacterial cell through efflux pumps [19]. Some bacteria prevent the antibiotic from reaching its target site by modifying the site such as resulting in reduced affinity to the antibiotic molecule [18]. Another common mechanism of bacterial resistance against antibiotics is the alteration in its target sites. This takes place as a result of a series of mutations occurring continuously in the presence of the antibiotic [20]. Circumvention of the susceptible metabolic pathway is another mechanism adopted by microorganisms for their survival [18, 21]. This leads to an overproduction of antibiotic targets resulting in resistance. Sometimes bacteria evolve in such a manner that they form new targets with similar functions as the original ones but are not susceptible to antibiotics [18]. This confuses the antibiotic to bind to the new target. Figure 1.1 shows the common mechanism of bacterial resistance against antibiotics. The treatment of antibiotic-resistant bacteria and the interpretation of susceptibility patterns may vary depending on the clinical history and the available treatment options [18].

1.4 Bioinformatics Strategies for Identification of Antibiotic Resistance Genes and Virulence Determinants

Bacterial resistance towards commonly used antibiotics raises serious health concerns worldwide, and relentless efforts are being made by the research community for its control. More recently, advancement in next generation sequencing (NGS) has revolutionized medical microbiology and shifted the phenotypic-based to genotypic-based identification and analysis of antimicrobial resistance. NGS can use DNA or RNA (as cDNA) for sequencing depending on the type of information required for downstream analysis. Sequencing of DNA can reveal the presence of antimicrobial resistance genes, whereas RNA-sequencing can help in the detection of gene expression, including the expression of antimicrobial resistance genes when coupled with appropriate bioinformatics pipelines, thus making it ideal for unraveling all possible genetic determinants of antimicrobial resistance in a single microbial genome.

1.4.1 Whole Genome Sequencing for Identification of Antibiotic Resistance Genes and Virulence Factors

Whole genome sequencing has enabled the rapid identification of resistance mechanisms, particularly in the context of drug-resistant tuberculosis (DR-TB), which is considered as a global health emergency [22, 23]. A combination of multiple antibiotics is given for TB treatment to minimize the chance of treatment failure due to the emergence of resistance in the course of treatment [24]. WGS has recently highlighted the cross-resistance of bedaquiline and clofazimine due to mutational upregulation of an efflux pump during the phase II trial [25]. WGS interrogates the entire genome; thus, it has become an important tool for identifying resistance during clinical trials and plays an important role in distinguishing exogenous reinfection from relapse of the primary infection. This is not possible when using traditional epidemiological tools which investigate only a little part of the genome [26–28]. WGS will soon become the gold standard for clinical trials of newer antibiotics against infectious diseases associated with recurrent disease [29, 30]. A study by Safi et al. identified a novel resistance mechanism against ethambutol as a result of the synonymous mutation in the gene *Rv3792*, which would have been excluded in the analysis for identification of the genetic basis of resistance [31]. In addition WGS also enables measuring the rate at which resistance emerges. This role has been highlighted by Ford et al. showing the elevated rate of acquiring rifampicin resistance in lineage 2 (Beijing) of MTBC (*Mycobacterium tuberculosis* complex) in comparison to lineage 4 (Euro-American lineage) [32–34]. This suggests that patients infected with lineage 2 isolates have a higher risk of developing multidrug resistance (resistance to isoniazid and rifampicin), thus highlighting the importance of early and active case detection [34].

With recent improvements in sequencing technologies, whole genome sequencing (WGS) has become an essential tool for controlling antibiotic resistance with the

Table 1.1 List of software's used for prediction of antibiotic resistance and virulence factors

Software's	Features of software	Reference
TypeWriter	Identifies antimicrobial resistance genes and SNPs from de novo assembly	[36]
Virulence Searcher	Predicts putative virulence factors from unannotated bacterial genomes	[37]
VirulentPred	Predicts bacterial virulent protein sequences using SVM approach	[38]
Bacterial Toxin Prediction Server (BTXpred)	Predicts bacterial toxins and their functions from primary amino acid sequences using SVM, HMM, and PSI-Blast	[39]
SPAAN	Prediction of adhesins and adhesin-like proteins using neural network	[40]
PhyResSE	Delineates MTB antibiotic resistance and lineage from WGS	[41]
Mykrobe	Offline species identification and prediction of antibiotic resistance in MTB from WGS	[42]
PointFinder	Detection of chromosomal point mutations associated with antimicrobial resistance from WGS by creating de Bruijn graph of contigs	[43]
ResFinder	Prediction of antibiotic resistance genes from both assembled and sequenced reads generated	[44]
ARIBA	Identifies antimicrobial resistance genes and SNPs from raw reads	[45]
KmerResistance	Identifies antimicrobial resistance genes and SNPs from raw reads based on k-mer	[46]
SRST2	Identifies antimicrobial resistance genes and SNPs from raw reads	[47]
GeneFinder	Identifies antimicrobial resistance genes and SNPs from raw reads	[36]
PARGT	A software tool for antibiotic resistance prediction by SVM	[48]
AMRFinder	A software tool for antibiotic resistance prediction by HMM	[49]
VAMPr	A software tool for antibiotic resistance prediction by machine learning approach	[50]

identification of resistant determinants. WGS is involved in various applications like the development of novel antibiotics and diagnostic tests. It also helps in elucidating the factors responsible for the emergence and persistence of bacterial resistance against antibiotics. The role of WGS as an important tool for controlling antibiotic resistance has been highlighted in many recent studies. WGS provides a flood of raw data, and harnessing this high-throughput information requires software to integrate it with omics data to get a wider overview of systems-level information. The first step after obtaining raw data from the sequencer involves the quality check which is done using FastQC and Trimmomatic tool. After quality control the data can be mapped to antimicrobial resistance databases as reads or assemblies of larger contigs. Various bioinformatics software (Table 1.1) are available which can process WGS raw reads or assembled contigs for characterization of resistance determinant

genes [35]. However, there is a vital need for the choice of an appropriate data analysis platform.

In the assembly-based method, the raw reads obtained from the sequencer are initially assembled *de novo* using different assemblers such as SPAdes, MEGAHIT, and Velvet followed by the prediction of protein-coding genes. Finally, the protein-coding genes are screened against the publicly available antimicrobial databases. When using assembly, the methods used for comparing input data with the AMR database are based on BLAST and hidden Markov model (HMM) searches. The specificity of results obtained depends on the criteria selected for gene length and percentage of similarity. However, the assembly-based method requires more computational expertise. Genome assembly is either reference-based or *de novo*. Reference-based assembly uses an existing reference template, and SNP detection is less accurate if the distance between isolate and reference is more. The *de novo* approach produces more fragmented genome assemblies and may miss the identification of antibiotic resistance genes if split across multiple contigs. It is also challenging to identify multicopy genes associated with antibiotic resistance in a *de novo* assembled genome. Sometimes a mutation in just one copy of the gene is sufficient to cause antibiotic resistance, but the assembly algorithm involves consensus of sequence; thus, significant genetic variants within repeats could be missed if repeats are collapsed into a single copy. This problem can be overcome by using long reads for assembly on increasing sequencing coverage, but that will increase the cost. Some of the software which take assemble data as input are tabulated in Table 1.1.

Read-based methods are also based on the alignment of raw reads to the AMR database but using different tools such as Bowtie2, BWA, and KMA [35]. KMA has been recently developed to map raw reads against redundant AMR database directly [51]. KMA was developed specifically for accurate bacterial genome analysis as BWA was being empirically used in microbiology but was originally developed for large reference genomes such as human genomes [51]. It uses k-mer seeding and Needleman-Wunsch algorithm for rapid and accurate alignment. The read-based method has an advantage over the assembly-based method as it can identify AMR genes present in low abundance which might be excluded from the analysis in case of incomplete assemblies [35]. There might be a chance of false-positive detection due to sequencing errors or DNA contamination in individual reads, but it can be reduced by setting the minimum threshold for the number of reads required for a positive outcome. Both the methods either read based or assembly based depend on the alignment to a panel of known antibiotic resistance genes and their associated mutations as curated in publicly available databases, thus, probably producing accurate results with high confidence in clinical settings. Unlike traditional antimicrobial resistance testing, WGS does not require several days for additional information but only approximately 24 h or less time to generate lot of data regarding drug susceptibility. Therefore, this reduces the time to start effective antibacterial therapy.

1.4.2 Antibiotic Resistance Database and Virulence Factor Database

Prediction of antibiotic resistance using WGS is an estimation of phenotypic antibiotic resistance as obtained from a culture-based antibiotic resistance test (gold standard). The susceptible and resistant strains are classified according to CLSI or EUCAST guidelines. [3] The accuracy of in silico AMR gene prediction depends on the selection of an appropriate AMR database irrespective of any bioinformatics approach used. Some AMR databases might consist of genes resistant to specific antimicrobials and specific bacterial species. Multispecies databases include CARD (comprehensive antibiotic resistance database) [52], ResFinder [44], PointFinder [43], ARG-ANNOT [53], and ARDB [54]. In addition, there are specific databases for *M. tuberculosis* such as TBDReaMDB [55] and MUBII-TB-DB [56]. The inclusion criteria of different databases also differ. For example, CARD contains entries published in scientific literature, whereas another database ResFinder does not have a strict publication requirement but consists of genes from GenBank. Most of the database includes only AMR genes, while very few of them include AMR genes along with their associated mutations. The format of entries and their possibility to download also vary in different databases. The user needs to decide which database fits their purpose for accuracy and specificity. The available databases with their features are listed in Table 1.2.

Identification of virulence genes in a genome utilizes a homology-based alignment to the virulence factors database. There are various databases having information related to virulence factors of different bacterial species as tabulated in Table 1.3.

Table 1.2 List of antibiotic resistance database

Database	Features of database	Reference
Antibiotic Resistance Genes Online (ARGO)	Contains 555 β -lactamases and 115 vancomycin resistance genes	[57]
Antibiotic Resistance Database (ARDB)	Contains 23,137 resistance genes from 1737 bacterial species	[54]
Repository of Antibiotic Resistance Cassettes (RAC)	Contains 389 resistance cassettes	[58]
Comprehensive Antibiotic Resistance Database (CARD)	Contains 4221 genes causing resistance to different antibiotic classes such as β -lactams, aminoglycosides, tetracyclines, rifampin, macrolides, fluoroquinolones, and sulfonamides and with antibiotic efflux	[52]
Antibiotic Resistance Gene (ARG-ANNOT)	Contains 1689 genes causing resistance to different antibiotic classes (β -lactams, aminoglycosides, fosfomycin, fluoroquinolones, glycopeptides, macrolide-lincosamide streptogramin, phenicols, rifampin, sulfonamides, tetracyclines, and trimethoprim)	[53]

Table 1.3 List of virulence factors database

Database	Features of database	Reference
Virulence Factor Database (VFDB)	Contains 29,746 virulence factor-related genes as well as anti-virulence genes of bacterial pathogens	[59]
Victors database	Contains 5173 virulence factors from 193 different human and animal pathogens	[60]
PathogenFinder	Web-server that differentiates virulent from non-virulent strains of bacterial pathogens using both protein and genome sequences	[61]
Toxin and Toxin Target Database (T3DB)	Comprehensive database of toxins and their targets containing 2900 peptide toxins, 1300 toxin targets and >33,000 toxin-target associations	[62]
Database of Bacterial Exotoxins for Human (DBETH)	A comprehensive database of human pathogenic bacterial toxins containing 229 toxins from 26 bacterial genus	[63]
VirmugenDB	Database of virulent genes used for the development of live attenuated vaccines	[64]
Pathogen-Host Interaction database (PHI-base)	Contains information on genes proven to affect the outcome of <i>pathogen-host</i> interactions	[65]
Pathosystems Resource Integration Center (PATRIC)	An online resource that stores and integrates genomic, proteomic, and transcriptomic data as well as protein-protein interactions	[66]
MvirDB	Contains information related to toxins, virulence factors, and antibiotic resistance genes	[67]

1.4.3 Machine Learning Tools to Predict Antibiotic Resistance and Virulence Factors

With the increase in the availability of genomic datasets and AST phenotypes, machine learning algorithms based on genotype data are known to be a promising tool for the identification of antimicrobial resistance. Machine learning uses either reference-based or reference-free algorithm to develop models correlating genotypic and phenotypic variations. This is a type of supervised learning which involves an input and a trained set based on the available information (expected outcome). The computational tools involve support vector machine (SVM) and set covering machine (SCM) for reference-based and reference-free algorithm, respectively [68]. The common strategy involves the training of a classifier based on important *ab initio* genetic features such as SNPs, indels, etc., or another strategy can be used by training on the basis of known important features retrieved from existing literature and database. A combination of both these strategies can also be used to curate a training set.

Machine learning tool was recently used in a study to predict antibiotic resistance in 16,668 *M. tuberculosis* isolates that have undergone drug-susceptibility testing (DST) for 14 antituberculous drugs [69]. Another study used artificial intelligence and machine learning-based prediction of resistant and susceptible mutations in *Mycobacterium tuberculosis* [70]. Moradigaravand et al. conducted a study on

1936 *E.coli* strains including existing and novel whole genome sequences highlighting the fact that machine learning approaches allow AMR gene prediction without prior knowledge of the mechanisms [71]. These approaches demonstrate that machine learning can robustly predict drug resistance with associated mutations and can be integrated in diagnostic tools. Machine learning has also found its role in predicting virulence factors in a bacterial genome. Several studies have been reported to use machine learning as a tool for virulence factor determination. A study on 115 clinical *P. aeruginosa* isolates (genotypically and phenotypically diverse) identified virulence determinants using genomic information applying a machine learning algorithm [72].

1.4.4 Identification of Pathogenicity Islands

Pathogenicity islands (PAIs) are genomic fragments that contain virulence genes and are horizontally transferred from other bacteria [73–75]. Detection of PAIs in a bacterial genome not only enables the identification of all virulence genes but also those that are not expressed by the bacterium. Several databases are presently available for the prediction of putative PAIs. The list of databases is tabulated in Table 1.4. These databases can be very useful for beginners, medical scientists, and microbiologists who are not familiar with computer languages. The other approaches involve the execution of computational commands under Linux systems. These approaches are either comparative genomics-based or sequence composition-based.

A comparative genome-based approach compares the genomes of closely related species which are assumed to have similar signatures. If a species has some special signatures which are not present in the closely related species, then those signatures are considered to be of foreign origin as a result of HGT. This approach involves three steps: (1) collecting all genome sequences from closely related species for a query genome; (2) aligning these genome sequences together; and (3) considering those gene segments present in the query genome but not present in others to be islands. This approach is advantageous in identifying the difference in signatures of closely related species but will be a problem for those species for which closely

Table 1.4 List of pathogenicity islands database

Database	Features of database	Reference
Pathogenicity island database (PAIDB)	Contains 223 types of PAIs predicted from 2673 prokaryotic genomes	[76]
The horizontal gene transfer database (HGT-DB)	Consists of 479 genomes with G+C content, codon and amino-acid usage information, as well as gene deviation in these parameters for prokaryotic complete genome	[77]
IslandViewer database	A database containing 18,919 virulence genes from 1277 bacterial pathogen genomes	[78]
PredictBias database	Contains virulence factors from 213 protein families	[79]
Islander database	Contains 3927 islands in 1302 genomes	[80]

related genomes are not available. Another disadvantage of this approach is a manual interpretation for adjustment and selection in most of the computational tools, leading to incompatible selection criteria due to unawareness of different genome structures [81].

Sequence composition-based approach compares genomic region within the same genome to identify special signatures such as G+C content, dinucleotide frequencies, codon usage, mobility genes, tRNA genes, and flanking direct repeats. The genomic regions within a genome share the same genomic signature; hence, if a part of genomic sequence detects different gene signatures, it is most probable that a particular region is horizontally transferred from foreign sources. This approach has an advantage over the comparative genomics approach, as it relies on the query genome and not genomes from closely related species, thus making it possible to predict genomic islands of all genomic sequences.

1.5 Limitations of In Silico Prediction of Virulence Factors and Antibiotic Resistance

In silico prediction of virulence factors and antibiotic resistance genes is a valuable strategy but has some drawbacks. The virulence factors might be present in non-pathogenic strains as well. Virulence might also depend on the expression level of virulence factors, and the mere presence of these factors might not confer virulence. This fact has been highlighted in a study showing 68.8% of 1988 virulence genes were present in both pathogens and non-pathogens [82]. It was observed in a study that *Pseudomonas aeruginosa* virulence varied in different hosts [83]. In *Neisseria meningitidis*, single-sequence repeat tracts and site-specific recombinations control virulence factors like capsular type, LPS structure, pilin diversity, and outer membrane protein expression [84]. WGS plays a very important role in predicting bacterial antibiotic resistance but has some limitations such as in the case of incomplete genomes, some of the integrons or transposons that play a role in antibiotic resistance might be present in the contig gaps and thus missed. WGS will not be able to predict the actual resistance determinant as resistance depends on several mechanisms. Specificity and sensitivity errors in genome prediction of antimicrobial phenotypes will lead to different consequences of treatment. The most concerning is false-negatives as it results in the inapt treatment of a resistant infection, thus increasing morbidity and fatality. Hence, reducing false-negatives at the cost of increasing false-positives is preferred. However, false-positives might give rise to last-line antibiotic resistance due to inappropriate use of antibiotics. To improve the accuracy of tools predicting WGS antimicrobial susceptibility, a large and diverse training dataset is required.

These limitations demonstrate that advancement in microbial informatics, particularly the development of more databases, using a combination of predictions by different tools, new bioinformatics algorithms, and protocols must be encouraged for specific, efficient, and accurate predictions keeping in mind the limitations encountered during analysis. Although there has been advancement in science with the

integration of bioinformatics, certain molecular characteristics such as virulence cannot be accurately predicted using *in silico* tools and need to be further validated *in vitro*.

1.6 Concluding Remarks

Antibiotics are increasingly being used for the treatment of bacterial infection giving rise to various resistant strains, thus threatening the efficacy of many antibiotics. Bacterial resistance can be prevented by a molecular understanding of the existence and activities of a bacterial pathogen. Recently bioinformatics has emerged as a vital tool for drug discovery, genomics, proteomics, and transcriptomics analysis to control bacterial infections as well as respond to epidemic outbreaks. The appropriate knowledge of the resistance mechanism combined with the bioinformatics approach can aid in predicting antimicrobial resistance and can successfully help improved therapeutic outcome.

References

1. Piddock LJV (2016) Reflecting on the final report of the O'Neill review on antimicrobial resistance. *Lancet Infect Dis* 16(7):767–768
2. Jorgensen JH, Ferraro MJ (2009) Antimicrobial susceptibility testing: a review of general principles and contemporary practices. *Clin Infect Dis Off Publ Infect Dis Soc Am* 49(11):1749–1755
3. Kahlmeter G (2014) Defining antibiotic resistance-towards international harmonization. *Ups J Med Sci* 119(2):78–86
4. Hu X, Xu B, Yang Y, Liu D, Yang M, Wang J et al (2013) A high throughput multiplex PCR assay for simultaneous detection of seven aminoglycoside-resistance genes in Enterobacteriaceae. *BMC Microbiol* 13:58
5. Bugarel M, Granier SA, Weill F-X, Fach P, Brisabois A (2011) A multiplex real-time PCR assay targeting virulence and resistance genes in *Salmonella enterica* serotype Typhimurium. *BMC Microbiol* 11:151
6. Punina NV, Makridakis NM, Remnev MA, Topunov AF (2015) Whole-genome sequencing targets drug-resistant bacterial infections. *Hum Genomics* 9:19
7. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J et al (2014) Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. *J Clin Microbiol* 52(4):1182–1191
8. Luscombe NM, Greenbaum D, Gerstein M (2001) What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med* 40(4):346–358
9. Saeb AT, Abouelhoda M, Selvaraju M, Althawadi SI, Mutabagani M, Adil M et al (2017) The use of next-generation sequencing in the identification of a fastidious pathogen: a lesson from a clinical setup. *Evol Bioinforma Online* 12:1176934316686072
10. Wu H-J, Wang AH-J, Jennings MP (2008) Discovery of virulence factors of pathogenic bacteria. *Curr Opin Chem Biol*. 12(1):93–101
11. Hacker J, Kaper JB (2000) Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* 54:641–679
12. Gal-Mor O, Finlay BB (2006) Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol* 8(11):1707–1719

13. Mahan MJ, Slauch JM, Mekalanos JJ (1993) Selection of bacterial virulence genes that are specifically induced in host tissues. *Science* 259(5095):686–688
14. Hensel M, Shea JE, Gleeson C, Jones MD, Dalton E, Holden DW (1995) Simultaneous identification of bacterial virulence genes by negative selection. *Science* 269(5222):400–403
15. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM et al (2014) Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J Clin Microbiol* 52(5):1501–1510
16. Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ et al (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409(6819):529–533
17. Hawkey PM (1998) The origins and molecular basis of antibiotic resistance. *BMJ* 317(7159):657–660
18. Munita JM, Arias CA (2016) Mechanisms of antibiotic resistance. *Microbiol Spectr* 4(2)
19. McMurry L, Petrucci REJ, Levy SB (1980) Active efflux of tetracycline encoded by four genetically different tetracycline resistance determinants in *Escherichia coli*. *Proc Natl Acad Sci U S A* 77(7):3974–3977
20. Lambert PA (2005) Bacterial resistance to antibiotics: modified target sites. *Adv Drug Deliv Rev* 57(10):1471–1485
21. Flensburg J, Sköld O (1987) Massive overproduction of dihydrofolate reductase in bacteria as a response to the use of trimethoprim. *Eur J Biochem* 162(3):473–476
22. Roemer T, Boone C (2013) Systems-level antimicrobial drug and drug synergy discovery. *Nat Chem Biol* 9(4):222–231
23. Zumla A, Nahid P, Cole ST (2013) Advances in the development of new tuberculosis drugs and treatment regimens. *Nat Rev Drug Discov* 12(5):388–404
24. Müller B, Borrell S, Rose G, Gagneux S (2013) The heterogeneous evolution of multidrug-resistant *Mycobacterium tuberculosis*. *Trends Genet TIG* 29(3):160–169
25. Hartkoorn RC, Uplekar S, Cole ST (2014) Cross-resistance between clofazimine and bedaquiline through upregulation of MmpL5 in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 58(5):2979–2981
26. Niemann S, Köser CU, Gagneux S, Plinke C, Homolka S, Bignell H et al (2009) Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. *PLoS One* 4(10):e7407
27. Eyre DW, Babakhani F, Griffiths D, Seddon J, Del Ojo Elias C, Gorbach SL et al (2014) Whole-genome sequencing demonstrates that fidaxomicin is superior to vancomycin for preventing reinfection and relapse of infection with *Clostridium difficile*. *J Infect Dis* 209(9):1446–1451
28. Dettman JR, Rodrigue N, Aaron SD, Kassen R (2013) Evolutionary genomics of epidemic and nonepidemic strains of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci U S A* 110(52):21065–21070
29. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P et al (2013) Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *Lancet Respir Med* 1(10):786–792
30. Menzies D (2013) Molecular methods for tuberculosis trials: time for whole-genome sequencing? *Lancet Respir Med* 1(10):759–761
31. Safi H, Lingaraju S, Amin A, Kim S, Jones M, Holmes M et al (2013) Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl- β -D-arabinose biosynthetic and utilization pathway genes. *Nat Genet* 45(10):1190–1197
32. Gagneux S, Small PM (2007) Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect Dis* 7(5):328–337
33. Hanekom M, Gey van Pittius NC, McEvoy C, Victor TC, Van Helden PD, Warren RM (2011) *Mycobacterium tuberculosis* Beijing genotype: a template for success. *Tuberc Edinb Scotl* 91(6):510–523
34. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T et al (2013) *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat Genet* 45(7):784–790

35. Boolchandani M, D'Souza AW, Dantas G (2019) Sequencing-based methods and resources to study antimicrobial resistance. *Nat Rev Genet* 20(6):356–370
36. Mason A, Foster D, Bradley P, Golubchik T, Doumith M, Gordon NC et al (2018) Accuracy of different bioinformatics methods in detecting antibiotic resistance and virulence factors from *Staphylococcus aureus* whole-genome sequences. *J Clin Microbiol* 56(9)
37. Underwood AP, Mulder A, Gharbia S, Green J (2005) Virulence Searcher: a tool for searching raw genome sequences from bacterial genomes for putative virulence factors. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis* 11(9):770–772
38. Garg A, Gupta D (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics* 9:62
39. Saha S, Raghava GPS (2007) BTXpred: prediction of bacterial toxins. *In Silico Biol* 7(4–5):405–412
40. Sachdeva G, Kumar K, Jain P, Ramachandran S (2005) SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinforma Oxf Engl*. 21(4):483–491
41. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM et al (2015) PhyResSE: a web tool delineating mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data. *J Clin Microbiol* 53(6):1908–1914
42. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B et al (2015) Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat Commun* 6:10063
43. Zankari E, Allesøe R, Joensen KG, Cavaco LM, Lund O, Aarestrup FM (2017) PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J Antimicrob Chemother* 72(10):2764–2768
44. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V et al (2020) ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother* 75(12):3491–3500
45. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA et al (2017) ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genomics* 3(10):e000131
46. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H et al (2014) Benchmarking of methods for genomic taxonomy. *J Clin Microbiol* 52(5):1529–1539
47. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T et al (2014) SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* 6(11):90
48. Chowdhury AS, Call DR, Broschat SLPARGT (2020) a software tool for predicting antimicrobial resistance in bacteria. *Sci Rep* 10(1):11033
49. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I et al (2019) Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob Agents Chemother* 63(11)
50. Kim J, Greenberg DE, Pifer R, Jiang S, Xiao G, Shelburne SA et al (2020) VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. *PLoS Comput Biol* 16(1):e1007511
51. Clausen PTL, Aarestrup FM, Lund O (2018) Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* 19(1):307
52. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A et al (2020) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* 48(D1):D517–D525
53. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L et al (2014) ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother* 58(1):212–220
54. Liu B, Pop M (2009) ARDB—antibiotic resistance genes database. *Nucleic Acids Res* 37(Database issue):D443–D447
55. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB (2009) Tuberculosis drug resistance mutation database. *PLoS Med* 6(2):e2

56. Flandrois J-P, Lina G, Dumitrescu O (2014) MUBII-TB-DB: a database of mutations associated with antibiotic resistance in *Mycobacterium tuberculosis*. *BMC Bioinformatics* 15:107
57. Scaria J, Chandramouli U, Verma SK (2005) Antibiotic Resistance Genes Online (ARGO): a Database on vancomycin and beta-lactam resistance genes. *Bioinformation* 1(1):5–7
58. Tsafnat G, Copty J, Partridge SR. RAC (2011) Repository of antibiotic resistance cassettes. *Database J Biol Databases Curation* 2011:bar054
59. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y et al (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* 33(Database issue):D325–D328
60. Sayers S, Li L, Ong E, Deng S, Fu G, Lin Y et al (2019) Victors: a web-based knowledge base of virulence factors in human and animal pathogens. *Nucleic Acids Res* 47(D1):D693–D700
61. Cosentino S, Voldby Larsen M, Møller Aarestrup F, Lund O (2013) PathogenFinde—distinguishing friend from foe using bacterial whole genome sequence data. *PLoS One* 8(10): e77302
62. Lim E, Pon A, Djoumbou Y, Knox C, Shrivastava S, Guo AC et al (2010) T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res* 38 (Database issue):D781–D786
63. Chakraborty A, Ghosh S, Chowdhary G, Maulik U, Chakrabarti S (2012) DBETH: a database of bacterial exotoxins for human. *Nucleic Acids Res* 40(Database issue):D615–D620
64. Racz R, Chung M, Xiang Z, He Y (2013) Systematic annotation and analysis of “virmugens”—virulence factors whose mutants can be used as live attenuated vaccines. *Vaccine* 31 (5):797–805
65. Winnenburg R, Urban M, Beacham A, Baldwin TK, Holland S, Lindeberg M et al (2008) PHI-base update: additions to the pathogen host interaction database. *Nucleic Acids Res* 36 (Database issue):D572–D576
66. Snyder EE, Kampanya N, Lu J, Nordberg EK, Karur HR, Shukla M et al (2007) PATRIC: the VBI PathoSystems Resource Integration Center. *Nucleic Acids Res* 35(Database issue):D401–D406
67. Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, Slezak T (2007) MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res* 35(Database issue):D391–D394
68. Cui G, Fang C, Han K (2012) Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics* 13 Suppl 7(Suppl 7):S5
69. Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R et al (2019) Machine learning predicts accurately mycobacterium tuberculosis drug resistance from whole genome sequencing data. *Front Genet* 10:922
70. Jamal S, Khubaib M, Gangwar R, Grover S, Grover A, Hasnain SE (2020) Artificial intelligence and machine learning based prediction of resistant and susceptible mutations in *Mycobacterium tuberculosis*. *Sci Rep.* 10(1):5487
71. Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L (2018) Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLoS Comput Biol* 14(12):e1006258
72. Pincus NB, Ozer EA, Allen JP, Nguyen M, Davis JJ, Winter DR et al (2020) A genome-based model to predict the virulence of *Pseudomonas aeruginosa* isolates. *mBio* 11(4)
73. Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 14(7):1394–1403
74. Chiapello H, Bourgain I, Sourivong F, Heuclin G, Gendrault-Jacquemard A, Petit M-A et al (2005) Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. *BMC Bioinformatics* 6:171
75. Chiapello H, Gendrault A, Caron C, Blum J, Petit M-A, El Karoui MMOSAIC (2008) an online database dedicated to the comparative genomics of bacterial strains at the intra-species level. *BMC Bioinformatics* 9:498
76. Yoon SH, Park Y-K, Kim JF (2015) PAIDB v2.0: exploration and analysis of pathogenicity and resistance islands. *Nucleic Acids Res* 43(Database issue):D624–D630

77. Garcia-Vallve S, Guzman E, Montero MA, Romeu A (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res* 31 (1):187–189
78. Bertelli C, Laird MR, Williams KP, Lau BY, Hoad G, Winsor GL et al (2017) IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res* 45(W1): W30–W35
79. Pundhir S, Vijayvargiya H, Kumar A (2008) PredictBias: a server for the identification of genomic and pathogenicity islands in prokaryotes. *In Silico Biol* 8(3–4):223–234
80. Hudson CM, Lau BY, Williams KP (2015) Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes. *Nucleic Acids Res* 43(Database issue):D48–D53
81. Ou H-Y, He X, Harrison EM, Kulasekara BR, Thani AB, Kadioglu A et al (2007) MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands. *Nucleic Acids Res* 35(Web Server issue):W97–W104
82. Niu C, Yu D, Wang Y, Ren H, Jin Y, Zhou W et al (2013) Common and pathogen-specific virulence factors are different in function and structure. *Virulence* 4(6):473–482
83. Dubern J-F, Cigana C, De Simone M, Lazenby J, Juhas M, Schwager S et al (2015) Integrated whole-genome screening for *Pseudomonas aeruginosa* virulence genes using multiple disease models reveals that pathogenicity is host specific. *Environ Microbiol* 17(11):4379–4393
84. Bidmos FA, Bayliss CD (2014) Genomic and global approaches to unravelling how hypermutable sequences influence bacterial pathogenesis. *Pathog Basel Switz* 3(1):164–184



Next Generation Sequencing: Opportunities and Challenges in Tuberculosis Research

2

Faraz Ahmad, Anwar Alam, Indu Kumari, Sugandha Singh,
Anshu Rani, Aquib Ehtram, Soumya Suhasini, Jasmine Samal, and
Nasreen Z. Ehtesham

Abstract

Despite a reasonable reduction in tuberculosis incidence and mortality in the last few years, the disease remains to be the major global public health challenge. The advent and transmission of drug-resistant tuberculosis (DR-TB) and delayed diagnosis and treatment have further worsened the control efforts against tuberculosis. The current culture- and PCR-based diagnostic methods are not sufficiently rapid and accurate enough to efficiently manage the patient load. While next generation molecular diagnostic assays such as the Xpert MTB/RIF and line probe assays (*MTBDRplus* and *MTBDRsl*) can only detect resistance-conferring mutations in specific target regions and miss novel emerging loci associated with resistance. Recent advancements in next generation sequencing (NGS) platforms and development of cost-efficient workflows have equipped the field with diagnostic tools that not only can detect existing mutations associated with drug resistance but can also identify newer coordinates of drug resistance. In a very short time frame, NGS provides whole genome sequence at a fairly low cost, that too without the need of culture (i.e., directly from the clinical specimen). NGS holds enormous potential as a tool for guiding personalized treatment through precision medicine as well as epidemiological monitoring of tuberculosis. In order to augment the accuracy and shorten the time taken to the accurate diagnosis of tuberculosis, NGS can become an unprecedented technology for diagnosis as well as epidemiological investigations of tuberculosis. In the present chapter, we will be discussing the available platforms, workflows, and tools based on NGS technology for the diagnosis and monitoring of tuberculosis.

F. Ahmad · A. Alam · I. Kumari · S. Singh · A. Rani · A. Ehtram · S. Suhasini · J. Samal ·
N. Z. Ehtesham (✉)

Inflammation Biology and Cell Signalling Laboratory, ICMR National Institute of Pathology,
Safdarjung Hospital Campus, New Delhi, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte
Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*,
https://doi.org/10.1007/978-981-16-0691-5_2

19

Keywords

Next generation sequencing (NGS) · Whole genome sequencing (WGS) · Tuberculosis; Diagnostics · Drug resistance monitoring · Molecular epidemiology · Single nucleotide polymorphism (SNP)

2.1 Introduction

The complete genome of MTB was deciphered and made available to the academic community in 1998 by Stewart Cole and colleagues at Institut Pasteur of Paris [1]. While decoding the genetic map of mono-chromosomal MTB genome, it was felt that mankind will soon mark a trump over this one of the most devastating and enigmatic pathogens in human history. However, the four-lettered genetic map encased in each MTB cell told us an otherwise strange tale of complex, reductive evolution of MTB for the last around 70,000 years [2, 3], making it best fit for survival in its favored host that's us, the humans.

At the time of annotation of the MTB genome, bacterial artificial chromosome (BAC) or cosmid/plasmid library-based shotgun sequencing approach was the method of choice and thus employed for the purpose. The shotgun sequencing method relies on Frederick Sanger and co-workers invented chain termination chemistry [4] that involves cloning of small (or large) DNA inserts in a suitable vector and creating a contiguous map of overlapping stretches to deduce and assemble the entire genome. However, it is a time- and labor-intensive effort that demands a great deal of infrastructural competencies and yet works at an unsatisfactorily slow pace.

With time, the need for improved sequencing methods was realized for rapid bacterial genomics applications. The pressing need for the search of the fast and efficient alternative to the chain termination-based sequencing methods was led to the development of next generation sequencing (NGS) technology for whole genome sequencing (WGS) applications by around 2005 with the introduction of the first commercial NGS machine, the GS20 (454 Life Sciences) [5]. GS20 enabled massively parallel sequencing of genome using a pyrosequencing approach, which involves sequencing of the solid-support (bead) immobilized DNA in an emulsion PCR. As the name suggests, NGS was supposed to overcome the limitations posed by the earlier methods that largely depend on the cloning of the genome required to be decoded. With the advent of the NGS platforms, the need for the cloning and maintenance of the target DNA has been alleviated. This leap in technology has endowed the field with unprecedented throughput (data output) capabilities and revolutionary reduction in the turnaround time (TAT) from several days to few hours and even real-time result availability in some of the most advanced instruments (Oxford Nanopore Technologies).

WGS technologies have been divided into three distinct developmental subcategories. (1) The first generation capillary electrophoresis (CE)-based sequencers that relied on Sanger's chain termination chemistry, for example,

3130xL (ABI/Life Technologies) and GeXP (Beckman Coulter). These CE-based sequencers were quite efficient with longer read length capabilities (up to 1000 bp) and reported to be minimally erroneous (error rate <0.1%). However, they were out-fashioned soon mainly because of their low-throughput power and relatively high cost. The shortcomings of first generation sequencers led to the development of (2), the second generation machines including 454 Genome Sequencer (GS) FLX, an improved version of GS20 (Roche¹), and HiSeq 2000/MiSeq (Illumina) that works on *sequencing by synthesis* technology or SOLiD (ABI) which works on the principle of *sequencing by ligation*. Second generation sequencer includes both bridge PCR-based platforms, for example, MiSeq and HiSeq 2500 (Illumina), and emulsion PCR-based semiconductor-enabled platforms such as Ion Torrent-personal genome machine (PGM) and ion proton (Thermo Scientific). The most recent advancement in the NGS field has been the introduction of (3), the so-called third generation platforms such as Sequel or RSII (Pacific Biosciences) and MinION or PromethION (Oxford Nanopore), which work on single-molecule real-time (SMRT) sequencing. The major technological advancement that third generation NGS platforms bring to the fore is the alleviation of the need for reference genome for assembly, as they generate longer reads (usually >1000 bp) and thus are capable of de novo genome assembly.

NGS-based WGS technology has found major application in the detection and surveillance of circulating drug-resistant (DR) isolates of dangerous pathogens. It helps implement both individual- and community-centric control measures for potential outbreaks caused by serious pathogens. WHO has identified NGS-based WGS as a major tool for the surveillance and control of antimicrobial resistance (AMR) priority pathogens, in addition to already prioritize pathogenic *Mycobacterium* species, including MTB [6]. A consortium of researchers, European Committee of Antimicrobial Susceptibility Testing (EUCAST), has already established an initial proof of concept regarding the development and harmonization of a universally agreed WGS-based antimicrobial susceptibility testing (AST) method for AMR priority pathogens including *Escherichia coli*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Salmonella* spp., *Klebsiella pneumoniae*, *Acinetobacter* spp., *Pseudomonas aeruginosa*, *Neisseria gonorrhoeae*, *Clostridium difficile*, etc., apart from MTB [7]. WGS can be implemented to monitor and to control nosocomial outbreaks involving near identical (genetically) strains that usually are unachievable using conventional methods [8].

In this chapter, we will be discussing the major advancements made possible due to the advent of NGS-based WGS technology in diagnosis and monitoring of susceptible and DR strains of MTB, vis-a-vis a brief discussion on available NGS technologies, their potential utility, and frequent challenges faced with their implication in tuberculosis research.

¹454 Life Sciences was acquired by Roche following the launch of GS20.

2.2 Application of NGS in Tuberculosis Research

2.2.1 Clinical Significance of NGS in Rapid Diagnosis of MTB and Its Drug Resistance Profiling

Although the genome of MTB was made available much before the introduction of NGS-based WGS technologies, it has outshined the earlier genome sequencing approaches by providing unparalleled depth, precision, and TAT while interrogating the mycobacterial and other genomes. NGS has become a powerful tool for rapid clinical diagnosis of drug-susceptible and drug-resistant TB, overcoming problems associated with the conventional diagnosis of TB and time taking (6–8 weeks) phenotypic drug susceptibility testing (DST) [9, 10].

2.2.1.1 Next Generation Diagnostics Applications

At the dawn of the twenty-first century, NGS has emerged as a revolutionary technique in the TB diagnostics field and enabled quick screening of the wide range of mutations in drug-resistant MTB strains. This has the potential to streamline a standard framework for clinical reporting of complete drug resistance profile in TB patients for successful treatment outcome. Based on NGS data, mutations conferring drug resistance can be classified as high, moderate, and low, for a given drug, as per their severity profile. WGS sequencing approach targets the complete genome of MTB at one time and identifies a wide range of specific nucleotide sequence changes to detect both the known and novel mutations associated with drug resistance. The genotypic resistance profile thus generated (predicted) can further be validated employing phenotypic DST for effective implementation into clinical settings [11].

2.2.1.2 Rapid DST Profile: Fast and Accurate Treatment Enabler

A recent study has documented the culture-free detection of MTB directly from sputum sample using NGS-enabled WGS and also obtained a drug resistance profile with an accuracy of 97.70% for the first-line drugs, when compared with conventional DST [12]. Similarly, sequencing on the Ion Torrent platform using “Ion AmpliSeq TB Research Panel” could rapidly identify novel variants and known mutations in the multiple drug resistance-conferring genes—*inhA*, *katG*, *rpoB*, *pncA*, *rpsL*, *embB*, *eis*, and *gyrA* [13]. NGS-based technology can also identify mixed infection along with MDR cases which will allow clinicians to formulate the drug regimen quickly for coinfection cases. Targeted genome sequencing for a selected set of drug resistance genes specific for pathogens can be achieved using NGS and can potentially be implemented in resource-limited settings for real-time monitoring of TB. It was reported that targeted amplicon sequencing is highly concordant with phenotypic DST, ranging from 94.6 to 98.8% for all the drugs [14].

2.2.1.3 Personalized Medicine Applications

Genetic variation of both the host and the pathogen should be taken into consideration while selecting the right choice of treatment for TB to prevent hepatotoxicity and unwanted drug toxicity. Wang et al. [15] have discovered that the isoniazid is

metabolized by N- acetyl transferase-2 (NAT-2) enzyme which catalyzes the acetylation step; the allele for NAT-2 enzyme is found to be highly polymorphic and thus categorized into fast and slow acetylator phenotype. In fast acetylation phenotype, the INH drug is easily metabolized and reaches the target site, whereas in the case of slow acetylation phenotype, the drug is metabolized slower and gets accumulated over the period to cause toxicity and drug resistance. Based on the INH-NAT2 model, NAT-2 genotype-guided isoniazid therapy has enormous potential in the prevention and management of TB [16].

Thus, NGS-enabled rapid DST provides unbiased rapid detection and characterization of drug resistance variants directly in the clinical sample to guide fast and accurate treatment. The individual resistance profile thus obtained opens the door for personalized medicine against TB based on personalized genomic information. A personalized medicine approach can be a game changer in curbing the tuberculosis epidemic, as compared to current standard therapy for DR forms of TB.

2.2.2 High-Throughput Screening to Detect Antibiotic Resistance and Novel Drug Targets of MTB

Over the time, WGS-based approaches have considerably advanced and aided in studying the existing laboratory-derived and circulating strains belonging to different lineages (lineage 1–7) among the diverse population that showed drug resistance of varying degrees to MTB. WGS studies have provided a whole account of single-nucleotide polymorphisms (SNPs) that are responsible for resistance to existing drugs in current treatment regimen and can help predict the possible mechanism of action of freshly discovered drugs [17] or the identification of the loci associated with resistance against newly introduced drugs such as bedaquiline, against which MTB is fast acquiring resistance.

2.2.2.1 Tracing Antibiotic Resistance Genes

There are many drug resistance databases of MTB based on WGS data. Various software have been developed to decipher drug resistance profile and lineage of bacteria from WGS raw data. Some of these tools are CASTB, KvarQ, Mykrobe Predictor, PhyResSE, Resistance Sniffer, TGS-TB, and TB Profiler, among others [18–23]. The input files for these tools are .fastq/.fq (single/paired-end files), .bam (binary alignment file), and .vcf (variant call file) files for analysis. These tools determine the level of resistance based on the earlier reported SNPs and their resistance to the antibiotics based on phenotypic testing.

2.2.2.2 Non-synonymous SNPs and Their Contribution to Resistance

Synonymous/silent mutations present in the genome generally don't alter the protein functions; therefore, non-synonymous mutations/SNPs are the focus of analysis in the WGS investigations [11]. Non-synonymous SNPs in drug interacting genes are largely the chief driver of the acquisition of drug resistance. The profile of drug resistance development shows the concerted evolution of resistance against multiple

drugs in MTB. For example, development of isoniazid resistance in MTB is always followed by sequential resistance development against rifampicin or ethambutol and pyrazinamide and further progressed to the resistance to next generation drugs [24]. The evolutionary trend of SNP acquisition also reflects the continuous micro-evolution occurring at the molecular level which also dictates the long-term evolutionary fitness of the mycobacteria. In addition to SNPs, insertion/deletion (InDel) variants too affect large or small sequence changes in the mycobacterial genome which can contribute to the development of resistance or virulence traits [7]. Therefore, InDels also represent an important point of interrogation while analyzing drug resistance coordinates in the bacterial genome and deserve considerable attention.

2.2.2.3 Virtual Screening of Novel Drug Targets

WGS analysis reveals potential drug targets against which candidate drug molecules can be designed and tested to obtain the potential drug leads. Drug designing approaches have been employed which involve the selection of the drug molecule based on structure-based drug designing, ligand-based drug designing, and virtual high-throughput screening of the ligands against the target protein/receptor [25]. The potential drug molecule obtained after virtual screening can further be optimized and validated using *in silico* tools, for example, by analyzing structure-activity relationship (SAR) property to better understand the events occurring at the molecular level and/or by assessing ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties of the candidate molecule [25]. Such virtual tools not only help select the best candidates with higher sensitivity, specificity, activity, and pharmacokinetics but also reduce the time, efforts, and cost expected to be exhausted in wet lab-based lead optimization and validation protocols. Furthermore, integrated approaches of combining *in silico* (virtual screening) and *in vitro* tools have the potential to accelerate the research and development of novel drugs against sensitive as well as DR (M-X-T-DR) strains of MTB for improved therapeutic management of tuberculosis.

2.2.3 Integrated NGS Approach to Better Understand the Transmission Dynamics and Evolution of MTB

2.2.3.1 Profiling of Novel Insertions and Deletions (InDels)

In a recent NGS-based WGS study of MTB isolates from north India, it was possible to detect more than 12,000 novel genetic variations comprising of 343 SNPs that were previously unknown in 38 genes. Various insertions of 2–67 bp were identified in the CDS region of the *pknA* gene, 5–34 bp in the *fadD34* gene, 2–30 bp in the oxidoreductase Rv0063 gene, 58 bp in genes associated with PE-PGRS proteins, and 16 bp in genes of PPE protein family, among others [26]. The SNPs aao-P240T, adh-H64L, and cons-L155S associated with D-amino acid oxidase, alcohol dehydrogenase, and cobalamin synthase, respectively, were found to be in higher proportion in DR isolates as compared to drug-sensitive isolates [26].

MTB is fast acquiring resistance to bedaquiline [27]. Mutations in the ATP synthase subunit C gene, *atpE*, are supposed to cause bedaquiline resistance [28]. Target-based assays could detect mutations in ATP synthase genes in only around 30% of the isolates. This pointed that there exist alternate mechanisms of resistance to bedaquiline. NGS analysis showed that SNPs associated with Rv0678 conferred resistance to bedaquiline. Although MTB MmpR5 (*Mycobacterium* membrane protein repressor 5) or Rv0678 protein does not interact with bedaquiline, it is associated with tolerance to both clofazimine and bedaquiline due to cross-resistance and also provides non-target-site-based resistance to bedaquiline. Mutations in Rv0678 gene cause upregulation of efflux pump protein MmpR5. Non-*atpE* mutants that are resistant to bedaquiline exhibit missense mutation (G281A) and single-nucleotide insertions in the Rv0678 gene [29, 30]. In another recent study, WGS analysis of more than 5000 MTB strains from around the globe showed that mutation in the *katG* gene (p.Ser315Thr) evolved earlier than *rpoB* gene mutations and other mutations conferring drug resistance across lineages, irrespective of geographic location [31]. Nearly 95% of the *rpoB* gene-mediated resistance to rifampicin occurs due to nucleotide deletions and insertions in the 81-bp hotspot region of RRDR (RIF resistance-determining region). Within the *rpoB* gene, point mutations in codons 531, 526, or 516 lead to a high degree of resistance to rifampicin, while mutations in codons 533, 522, 518, or 511 induce a low degree of resistance to rifampicin [32]. Mutations in the *katG* gene (S315T) or the addition of arginine in the 81-bp hotspot region of RRDR induces resistance to rifampicin.

2.2.3.2 Epigenomic Alterations

Epigenetic changes such as methylation and acetylation can determine the transcriptome profile of MTB. Epigenetic changes can modulate the expression of the gene by modifying interaction of transcription factors with DNA. Lineage-specific methylation patterns occur in several MTB lineages. Methylation sites may overlap with DNA sequences that bind to sigma factors, considered crucial for MTB pathogenesis. Bisulfite sequencing involves the use of sodium bisulfite to convert unmethylated cytosine into uracil and provides a high-resolution quantitative estimation of genome methylation. NGS mapping of methylated fragments provides an estimation of methylation at each CpG locus.

Studies using the single-molecule real-time (SMRT) sequencing method have illustrated mechanisms of 5-methylcytosine and N6-methyladenine-mediated methylation in MTB genomes [33]. The motifs within DNA methyltransferases (MTases), *hsdM*, *mamA*, and *mamB* modulate N6-methyladenine modification [33]. A loss of *mamA* MTase activity in MTB causes suppression of gene expression involved in survival during hypoxia [34].

2.2.3.3 Metagenomic Profiling

Metagenomic next-generation sequencing (mNGS) offers the advantage of requiring fewer amounts of sample, unbiased detection, and shortened TAT. mNGS enables the detection of pathogenic bacteria where patients do not respond to antitubercular therapy. For slow-growing bacteria, mNGS is a viable option as it reduces the

diagnosis time for confirmation of bacterial/fungal infection and hence speeds up the process of initiation of antimicrobial therapy and thus improves treatment prognosis [35]. mNGS is recommended in hospitals with limited detection tools for immunocompromised patients and those with complicated or severe infections. However, mNGS must always be substantiated with clinical symptoms and epidemiology database before final identification of pathogenic microbe.

A major disadvantage for the detection of TB using mNGS method is that MTB-specific nucleic acids are seldom present in the extracellular milieu because of it being an intracellular, obligate pathogen. Although mNGS can well distinguish MTB and nontuberculous mycobacteria (NTM), however, detection of particular species within the MTB complex can be affected when coverage is insufficient. In such a scenario, targeted PCR should be carried out to identify the specific MTB complex organism. The diagnostic sensitivity of mNGS is reduced in patients undergoing TB treatment due to either contamination of MTB genome during library preparation, poor-quality reads of high-complexity sequences, errors from database entries containing human DNA reads, misannotated species, and sequencing adaptors or arose due to introduction of contamination during sample collection and processing stages [36, 37].

2.2.3.4 Microevolution

The MTB genome exhibits a high degree of heterogeneity and less stability within the host. MTB can coevolve within the host to acquire resistance and develop into distinct subpopulations. Trauner et al. [38] noticed high levels of genetic divergence during microevolution and concluded that drug pressure and immune surveillance within the host modulate the stability of SNPs. Distinguishing the mixed type of infections and microevolution of strains that result due to clonal variation was a huge challenge. MTB cells present within the lung lesions can microevolve through acquisition of SNPs, which may explain spatiotemporal differences observed in MIC values in the same patient [39]. The proportion of bacilli strains in each lesion differs, and hence different variants of subpopulations may be detected in the sputum. MTB strains evolve initially in the lungs and disseminate to extrapulmonary organs where they evolve further. In this scenario, samples from different tissue exhibit distinct subpopulations of MTB. A TB patient may get superinfected with a phylogenetically similar strain, and the new evolved strain can be detected in the sampling process.

NGS technology has provided the impetus to not only establish the mycobacterial diversity in the mixed populations but also enables the clinicians to establish a chronology of mutations leading to M/X-DR. Retrospective WGS analysis of MTB infection in XDR-TB patients has successfully revealed the sequence of acquisition of mutations that led to the development of XDR including resistance to new drugs such as delamanid and bedaquiline [40]. NGS analysis of mutations (for INH and RIF) due to selection pressure on drug resistance-conferring loci over a period of 1 year demonstrated the sequential acquisition of MDR-TB [41].

An association between disease severity and genetic diversification of the pathogen is also reported. A majority of the loci that exhibit extreme variation are found in

genes conferring drug resistance and cell wall lipid transporters, which allow the pathogen to adapt within the host and aid in their dissemination. While monitoring the transmission chain, the SNP distance between isolates must be considered. A distance of less than 12 SNPs between any two strains indicates genetic linkage and direct transmission [42]. The mutation rate per year reported in the loci conferring drug resistance is seven SNPs per genome as compared to the rest of the genome where it is 1.1 SNPs per genome. The mutation rates per year show variation, from 0.3 to 7.0 SNPs per genome per year, between different strains [42]. However, some reports also suggest that some strains do not show any SNP difference even when there is a high level of diversity within-host [43]. The relapse and reinfection can also be distinguished by analyzing SNP distance. An SNP distance of less than 9 indicates relapse, while an SNP distance above 100 indicates reinfection [44]. Variants in the *pks5* and *fadD* gene families that regulate cell wall biosynthesis evolve within the host and confer drug resistance or aid in fitness compensation [45]. Mutations in *pks5* lead to INH resistance and are associated with the triggering of compensatory mechanisms. MDR strains of MTB acquire mutations in the Rv3303c, Rv2071, and Rv0888 genes in patients while undergoing TB treatment [46].

2.2.4 Epidemiological Surveillance to Combat Drug-Resistant Tuberculosis

2.2.4.1 Variants Calling and the Relation Between Genotype-Phenotype in Resistant Isolates

A clear understanding of the variant interpretation and correlation of genotype-phenotype will open the door of personalized medicine for a successful treatment outcome. Variant calling involves the identification of SNPs and small InDels from NGS data. Targeted NGS can be implemented in low-middle-income countries (LMICs) for drug resistance surveillance effectively [47]. It involves direct sequencing of drug resistance-associated known loci in the MTB genome from either clinical specimen (sputum) or primary liquid culture (MGIT) without the need for WGS and lengthy LJ culture [47, 48]. The obtained results can be aligned to some curated databases such as Relational Sequencing TB (ReSeqTB) (<https://platform.reseqtb.org/>), Sprint-TB (<http://www.sprinttb.org/>), and Pathgenseq (<http://pathgenseq.lshtm.ac.uk>) for the variant calling of resistance-conferring bases. NGS technology can detect multiple variants (previously uncharacterized) of drug resistance which can further be characterized and validated through phenotypic DST. In contrast, previously reported variants would be confirmed by variant call in an available database like ReSeqTB, to get an advanced genotypic DST result within fast TAT, enabling quick treatment decision, and the results can further be validated with phenotypic DST [49].

2.2.4.2 GWAS to Identify the Mutations in Drug-Resistance Associated Genes

Originally developed for interrogating recombination enabled dimorphic genomes, genome-wide association studies (GWAS) now have been repurposed to search for additional signatures of resistance in the monomorphic MTB genome. One such study involving next-gen WGS-based GWAS investigations on 123 MTB genomes identified resistance hotspots with 100% confidence and signatures for convergent evolution in 47 DR isolates [50]. The study revealed the genomic diversity based on SNP (non-synonymous) profiling and also discovered new genetic determinants associated with the drug resistance in MTB, particularly a novel mutation in the *ponA1* gene, conferring resistance to rifampicin [50]. GWAS convergence-based approach also identified the lineage-specific evolution of drug resistance [51]. It was further revealed that the shortest SNP difference between clinical isolates will be epidemiologically linked to represent transmission cluster [52]. Clustering of SNPs based on phylogenetic tree analysis of whole genome provides a better understanding of transmission dynamics as well as inter and intra-patient variations during an outbreak [42, 53, 54]. It also helps delineate whether the drug resistance originated in a single cluster is either due to transmission or innate acquisition of a particular drug resistance variant in a gene.

2.2.4.3 Genealogy of Recent Transmission Through Phylogenetic Tree Analysis

Advancements in the field of NGS-based WGS has enabled the reconstruction of the phylogenetic tree and dating the evolution of DR-TB to confirm the evolution and transmission of new clones. A study from Argentina has undertaken a phylogenetic analysis of WGS data from 252 isolates to track the origin and evolution of an outbreak associated XDR strain of MTB and deduced the timeline spanning four decades of circulation and evolution of the founder “M” strain, before the outbreak [55]. Clearly, NGS-based WGS methods not only deliver insights guiding appropriate clinical management of DR-TB but also provides an opportunity to track down the chain of ongoing transmission in order to avert future outbreaks.

Recently discovered anti-TB drugs bedaquiline and delamanid are the last resort option for treatment of MDR-TB and introduced in only some countries [56]. However, a resistance-associated mutation in bedaquiline and delamanid target genes has been identified in Tibetan patients with TB using WGS [57]. These resistant strains may fast disseminate in other regions, so it is crucial to follow the transmission of such resistant strains in order to check and halt the ongoing transmission [58].

2.3 Challenges for Implementation of NGS in Clinics and Mycobacterial Research Laboratories

NGS-based protocols and bioinformatics tools have been developed and being further optimized so that the time for the detection and characterization of DR-TB can be significantly reduced. The potential impact of NGS on disease diagnosis,

clinical decision-making, and public health can be revolutionary in terms of pace as well as volume. However, there are several shortfalls associated with the use of NGS which hampers the diagnosis and control of DR-TB in high-burden areas. The lack of specialized personnel to handle and analyze big datasets generated out of the NGS exercise is the major bottleneck here. We are looking at the challenges for the implementation of NGS for routine clinical use and also propose how these might be overcome basically in resource-limited, high TB burden regions.

2.3.1 Challenges Associated with Sample Preparation and Scale-Up

2.3.1.1 DNA Extraction Methods and Quality Control Issues

Different NGS workflows/platforms have different requirements regarding the quality and quantity of extracted DNA which always needs to be pre-determined. Some major issues such as degradation of primers due to repetitive freeze-thawing, inadequate amount of DNA template, and the inadequate removal of inhibitory impurities including ethanol, salts, phenol, etc. could result in bad or insufficient sequencing data [59]. Before proceeding with DNA extraction, one should refer to and decide the method of library preparation or kit being used as the kits available commercially generally demands minimum input DNA quantities for library preparation [59, 60].

2.3.1.2 The Culture Requirement for NGS of MTB and Other Associated Issues

Many factors influence the need for culturing the MTB for NGS-based applications that includes sequencing platform, library preparation method, desired coverage, or depth required for post-sequencing analysis, among others. Extraction of genomic DNA basically employs traditional methods of growing mycobacteria in the liquid (7H9 broth-based MGIT) or solid medium (Lowenstein Jensen (LJ) slants). The most important consideration while selecting the method for growing MTB is the required quantity and quality of the DNA for downstream applications. For routine WGS-based diagnostic purposes, DNA from the clinical specimen such as sputum would mostly be enough to run NGS. However, to obtain deeper insights for novel alterations in the genome, high-quality and quantity of DNA are a prime requirement. Solid media, such as LJ slants, ensure selective growth of MTB and provide higher DNA yields for sequencing purposes. However, this method increases the TAT for NGS.

Some recent studies have tried to address these challenges by refining existing methods. Votintseva et al. have re-optimized established mycobacterial DNA isolation methods to minimize contaminating DNA from early positive MGIT cultures. It enabled the extraction of MTB DNA above critical limits (≥ 0.2 ng/ μ l) for Nextera XT (Illumina)-based library preparation and subsequent identification of MTB with 98% concordance to reference laboratory using Illumina MiSeq NGS platform [61]. In another study, a genotypic DST profile was obtained from sputum MTB samples within 5 days of receipt using SureSelect XT (Agilent)-based targeted DNA enrichment strategy for 74% of samples tested. The method gave the power to obtain

DST results well before 24 and 31 days before the availability of MGIT culture-based WGS and phenotypic DST, respectively [48].

2.3.2 Available NGS Platforms: Selecting the Best for the Need

All the currently available commercial NGS platforms can be categorized majorly on the basis of their sequence length read abilities, either short-read- (Illumina) or long-read-enabled platforms (Oxford Nanopore); detection through amplification (Illumina Platforms) or real-time single-molecule detection (Sequel, Pacific Biosciences); or method of detection and the chemistry involved in the library preparation. Brief information about some of the common NGS platforms is mentioned in the following subsection.

2.3.2.1 Illumina Sequencers

Illumina is the biggest market player in the NGS field as of now and offers a range of low- to high-throughput capability platforms. Illumina-made sequencers are widely used platforms for NGS-based WGS and other applications. There are several instruments in the Illumina catalogue such as MiniSeq, MiSeq, HiSeq, NovaSeq, and NextSeq, each of them has different sample throughput and genome coverage. These instruments use a fluorescent-based detection system and also work on *sequencing by synthesis* approach [62]. Among all the above instruments, MiniSeq and MiSeq are most frequently used because they are economic as well as user-friendly [42, 63, 64].

2.3.2.2 Ion Torrent Machines

This instrument is manufactured by Thermo Fisher Scientific. This platform also employs *sequencing by synthesis* method, but the detection is based on semiconductor technology that measures the hydrogen ion concentration released during DNA polymerization by using solid-state pH sensor [62, 65]. This platform uses emulsion PCR chemistry for library preparation. The use of these complex techniques makes this platform a little expensive. Instruments under the Ion Torrent platform are PGM, S5, and Proton.

2.3.2.3 Pacific Biosciences (PacBio) Sequencers

The NGS platforms from Pacific Biosciences are significantly different from all the above mentioned machines. Although it employs the same fluorescent-based detection method, the library preparation does not include any amplification step [65, 66]. This platform works on the real-time SMRT technology (discussed in Sect. 2.1) sequencing technology. Instruments under this platform are PacBio RSII and Sequel. Sequel is the modified version of RSII having a high-throughput capability.

Table 2.1 A brief account on various NGS instruments available under all four above mentioned platforms and their comparative characteristics

Platforms/ instruments	Sequencing run time (approx.)	Instrument cost (USD) [65]	Throughput range (Gb) [67]	Read length (bp) [65]
<i>Illumina platforms</i>				
(a) MiniSeq	4–24 h [65]	50,000	1.7–7.5	2 × 150
(b) MiSeq	27 h [68]	100,000	0.3–1.5	2 × 300
(c) HiSeq	11 days [68]	650,000	10–1000	2 × 150
(d) NovaSeq	16–45 h [65]	250,000	2000–6000	2 × 150
(e) NextSeq	12–30 h [65]	850,000–950,000	10–120	2 × 150
<i>Ion Torrent (Thermo Fisher)</i>				
(a) PGM	2 h [68]	80,000	0.08–2	400
(b) S5	~19 h [65]	60,000	0.6–15	400
(c) Proton	4–24 h [65]	149,000	10–15	200
<i>Pacific Biosciences</i>				
(a) PacBioRSII	2 h [68]	750,000	0.5–1	60,000
(b) Sequel	~20 h [65]	350,000	5–10	60,000
<i>Oxford Nanopore</i>				
(a) MinION	30 min–48 h [65]	1000	10–20	100,000+
(b) GridION		2400	50–100	100,000+
(c) PromethION		25,000	480–960	100,000+

2.3.2.4 Oxford Nanopore Platforms

These platforms use biological nanopore technology for sequencing. It measures the change in electric conductivity when slightly differing molecular weight nucleotides (A, T, G, C) of DNA pass through a biological nanopore [62, 66]. The detector records and identifies each of the four nucleotide bases on the basis of slight alterations in conductivity and uses this information to deduce the sequence. This platform is highly portable (even pocket-sized) and the simplest among all. The available instruments under the Oxford Nanopore category are MinION, GridION, and PromethION (Table 2.1).

2.3.3 Selection Criteria

In a routine public health laboratory, required NGS essentialities are easy-to-use workflow, affordable setup, low per-sequence cost, intermediate throughput range, short run time, and portable system. Considering these preconditions, MiSeq, MiniSeq, PGM, and MinION are suitable options [67]. PGM instrument is somewhere similar to MiSeq, but the per-sample cost is higher. However, NextSeq, NovaSeq, and HiSeq instruments provide high-throughput range and low per-sample cost, but still are not suitable for endemic settings as they require extended automation for library preparation [66, 67].

2.3.4 Output Data Quality

The error rates of short-read platforms like Illumina are comparatively lower and hence are beneficial for infectious disease applications such as tuberculosis diagnosis and epidemiological investigations. Ion Torrent platforms possess higher error rates because of their long-read property, and moreover, the data output of Ion Torrent is also low as compared to Illumina. In GC-rich regions which are abundant in the MTB genome, the error rate by MiSeq instrument is less as compared to Ion Torrent and PacBio platforms [65–67].

2.3.5 Turnaround Time (TAT)

Most of the next generation sequencers are more or less similar in TAT that generally doesn't exceed more than a day. However, Oxford Nanopore Technologies-based platforms are quite time-efficient as compared to all other rivals providing real-time results and monitoring abilities. A comparative chart of various available NGS modules and their respective run time is presented in Table 2.2.

2.3.6 Handling a Large Amount of NGS Data: Frequent Challenges

2.3.6.1 Difficulties in Analysis

NGS has been established as a promising approach to understand genome organization, detection, and identification of SNPs responsible for drug resistance in MTB. However, frequent challenges are faced while handling and preparing the samples for sequencing and lack of standardized pipelines to analyze the raw data of sequencing. In the case of MTB, the problem which arises during sample preparation is due to its cell wall which is rigid, rich in lipids, and resistant to a number of lysis buffers [69]. The disruption of the cell wall affects the quantity and quality of DNA. There is one challenge which is related to the genome of MTB, i.e., the presence of repetitive/disordered regions and hard-to-sequence regions with high GC content. It can be overcome by increased genome-wide sequencing depth to sequence more efficiently and accurately or do long-read sequencing instead of short-read sequencing [70] or maybe a combination of both, short-read- (Illumina) and long-read (MinION, Oxford Nanopore)-enabled methods as reported recently with improved coverage (99.7%) and minimum omission [71].

2.3.6.2 Archival Issues

The bioinformatics analysis of MTB WGS data poses different problems like the adapters used while sequencing. The WGS data may remain in the raw form (.fastq files) and the presence of contaminant sometimes compromises the quality of the data and results. The developed pipelines need to be defined and optimized based on the disease-causing pathogens that range from prokaryotes to eukaryotes, haploid to diploid, and unicellular to multicellular organisms. As MTB is a prokaryote bacterial

Table 2.2 A comparative account on various currently available NGS platforms and their major advantages and limitations

Platform	Input DNA amount	Instrument	Advantages	Limitations
Illumina platforms	50–1000 ng [68]	MiniSeq	Run time is short, initial investment is low [65]	Read length is small, low-throughput range [67]
		MiSeq	Long read length [67]	Sequence run time is long [65]
		HiSeq	High-throughput, read accuracy [65]	Sequence run time is long, initial cost is high [67]
		NovaSeq	High-throughput, read accuracy [65]	Sequence run time is long, and initial cost is high [65]
		NextSeq	High-throughput [67]	Sequence run time is long [67]
Ion Torrent (Thermo Fisher)	100–1000 ng [68]	PGM	Short run time, long read length [65]	Low-throughput range, homopolymers [67]
		S5	Long read length [65]	Homopolymers [67]
		Proton	Short run time, long read length [67]	Homopolymers [67]
Pacific Biosciences	~1000 ng [68]	PacBio RSII Sequel	Short run time, long read length [67]	Error rate and initial cost are high [65]
Oxford Nanopore	400–1000 ng [68]	MinION	Short run time, portable, long read length, low-cost [65, 67]	Error rate is high [65]
		GridION	Short run time, long read length [65, 67]	
		PromethION	Short run time, long read length [65, 67]	

pathogen, there is a dire necessity to develop bioinformatics pipelines dedicated to it, which can be implemented for the analysis and interpretation of NGS data. During post-sequencing analysis, there comes a step of recalibration for the obtained variants in the haplotype caller. Therefore, there is a need for the presence of a standardized reference .vcf file that can be incorporated in any of the annotation databases. The available online tools give the results about the resistance of the sample; however, there are only few which describe the lineage of the strain. It is difficult to work with large amount of data with online tools because of data processing requirements. Therefore, there is an urge to develop standard offline protocol which can be easy to understand and execute. WGS technique is generating big data which keeps on increasing; therefore, it is a demand of time to analyze it in order to understand the evolving mechanisms of drug resistance, evolution, and transmission within the host or in the environment.

2.3.7 Evolving Knowledge on MTB Genome: Where We Are, and Where Do We Need to Be?

WGS data help to record the changes occurring at the gene level that can be utilized to understand the evolutionary aspect of slowly evolving MTB. The estimation of the evolutionary rate of MTB has been assessed by molecular clock utilizing phylodynamic models to characterize TB epidemics by determining the beginning and an end of the outbreak, ascertaining the origin and spread of resistant strains, correlating the population dynamics, and estimating the transmission dynamics over the time and geographical regions [72]. Kühnert et al. [73] have carried out a phylodynamic analysis to investigate and compare the epidemiological dynamics of two MTB outbreaks. It was difficult to calculate the rate of evolution on outbreak time scales [73]. The probability of the possession of resistance by normal rate of mutation is 1 in 10^8 bacilli for rifampicin, to ~ 1 in 10^6 bacilli for isoniazid, streptomycin, and ethambutol [24]. The available MTBC genomic data has shown five human-adapted lineages representing *M. tuberculosis sensu stricto* (L1–L4 and L7), two other lineages that belong to *M. africanum* (L5–L6), and nine lineages adapted to animals. *M. canettii*, which is restricted to the Horn of Africa, is related to earlier evolved lineages. It is a well-known fact that MTBC strains show a difference of ~ 2000 SNPs; *M. canettii* has 10- to 25-fold variations and is segregated by $\sim 14,000$ SNPs from the most recent common ancestor (MRCA) of MTB. Ngabonziza et al. [74] reported lineage L8 sister clade by genome-based phylogenetic reconstruction to the known MTBC lineages. This lineage has diverged by the loss of the *cobF* genome region that encodes for precorrin-6A synthase required for the biosynthesis of cobalamin/vitamin B12 [74]. It is documented that the rate of variation leading to drug resistance alters as per the lineage and strain. In vitro analysis has shown the association of DR-TB with increased mutations in the case of the Beijing strain family. Transmission dynamics of MTB has been studied by molecular genotyping tools like occurrence/removal of spacer sequences (spoligotyping), analysis of tandem repeat patterns (MIRU-VNTR), and currently WGS data. Spoligotyping can estimate the events that occurred 200 years back and MIRU-VNTR up to 30 years, and SNPs obtained from WGS can determine transmission events of timeline up to 10 years (5SNP/allele cutoff) [52]. It shows that the estimation of transmission dynamics depends on the number of nucleotide/sequences/blocks analyzed in the study. Therefore, to understand the transmission dynamics across different geographical time scale and regions, population, and lineages, the methodology should comprise of at least two molecular genotyping tools. MTB diversity directly evolved within-host was estimated by WGS, and it was reported that rare variants present in the genome are not fixed which leads to heterogeneity [75]. During transmission, the rare variants are sowed, and in the latency period, these can get selected. To establish recent transmission links, it is suggested that the minor alleles should be incorporated along with fixed SNPs. For future directions, the molecular clock and transmission dynamics of MTB are not well understood; therefore, there is a need to study the events which lead to the

development of more robust resistant strains against the existing drugs and newly developed drugs.

2.4 Conclusions

Harnessing the utility of NGS within the present diagnostic framework for both drug-susceptible and drug-resistant TB requires a standardized workflow that is widely available and congruous. In order to successfully explore the full potential of the NGS-based WGS technique for surveillance and clinical management of DR-TB, more intense efforts are needed to be invested. These efforts should aim to fully unravel its economic benefits and potential shortcomings in both low and high TB burden settings, as well as to build and add more authentications to the growing list of potential solutions to implementation challenges in the area. The potential adjustments should aim to upgrade existing capabilities in both laboratory infrastructure and key expertise such as bioinformatics, databases, and software development, which provide support and enable proper handling and translation of the massive amounts of big data generated through NGS into clinical reality. NGS-based WGS for MTB, especially direct from sputum, will play a pivotal role in the near future to stop developing the epidemic of TB, as this remarkably shortens the TAT to results and enables the provision of an effective treatment regimen to TB patients who might be harboring DR strains.

References

1. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D et al (1998) Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. *Nature* 393 (6685):537–544
2. Gagneux S (2012) Host-pathogen coevolution in human tuberculosis. *Philos Trans R Soc B Biol Sci* 367(1590):850–859
3. Ahmed N, Dobrindt U, Hacker J, Hasnain SE (2008) Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. *Nat Rev Microbiol* 6(5):387–394
4. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci [Internet]* 74(12):5463–5467. Available from <http://www.ncbi.nlm.nih.gov/pubmed/431765>
5. Goldberg B, Sichtig H, Geyer C, Ledebner N, Weinstock GM (2015) Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *MBio* 6(6):1–10
6. WHO (2017) Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics [Internet]. Available from https://www.who.int/medicines/publications/WHO-PPL-Short_Summary_25Feb-ET_NM_WHO.pdf?ua=1
7. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C et al (2017) The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clin Microbiol Infect* 23(1):2–22
8. Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, van Schaik W, Wertheim HFL (2017) Whole-genome sequencing of bacterial pathogens : the future of nosocomial. *Clin Microbiol Rev* 30(4):1015–1064

9. World Health Organization (2008) Policy guidance on drug-susceptibility testing (DST) of second-line antituberculosis drugs [Internet]. Available from WHO, 2008 (WHO/HTM/TB/2008.402)
10. Rodrigues C, Jani J, Shenai S, Thakkar P, Siddiqi S, Mehta A (2008) Drug susceptibility testing of Mycobacterium tuberculosis against second-line drugs using the Bactec MGIT 960 System. *Int J Tuberc Lung Dis* 12:1449–1455
11. Papaventsis D, Casali N, Kontsevaya I, Drobniewski F, Cirillo DM, Nikolayevskyy V (2017) Whole genome sequencing of Mycobacterium tuberculosis for detection of drug resistance: a systematic review. *Clin Microbiol Infect* [Internet] 23(2):61–8. Available from <https://doi.org/10.1016/j.cmi.2016.09.008>
12. Soundararajan L, Kambli P, Priyadarshini S, Let B, Murugan S, Iravatham C et al (2020) Whole genome enrichment approach for rapid detection of Mycobacterium tuberculosis and drug resistance-associated mutations from direct sputum sequencing. *Tuberculosis* [Internet] 121 (Feb):101915. Available from <https://doi.org/10.1016/j.tube.2020.101915>
13. Park J, Shin SY, Kim K, Park K, Shin S, Ihm C (2018) Determining genotypic drug resistance by ion semiconductor sequencing with the ion ampliSeq™ TB panel in multidrug-resistant mycobacterium tuberculosis isolates. *Ann Lab Med* 38(4):316–323
14. Colman RE, Anderson J, Lemmer D, Lehmkuhl E, Georghiou SB, Heaton H et al (2016) Rapid drug susceptibility testing of drug-resistant mycobacterium tuberculosis isolates directly from clinical samples by use of amplicon sequencing: a proof-of-concept study. *J Clin Microbiol* 54 (8):2058–2067
15. Wang P, Pradhan K, Zhong X-B, Ma X (2016) Isoniazid metabolism and hepatotoxicity. *Acta Pharm Sin B* [Internet] 6(5):384–392. Available from <https://doi.org/10.1016/j.apsb.2016.07.014>
16. Khan N, Das A (2020) Can the personalized medicine approach contribute in controlling tuberculosis in general and India in particular? *Precis Clin Med* 3(3):240–243
17. Cohen KA, Manson AL, Desjardins CA, Abeel T, Earl AM (2019) Deciphering drug resistance in Mycobacterium tuberculosis using whole-genome sequencing : progress, promise, and challenges. *Genome Med* 11(45):1–18
18. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM et al (2015) PhyResSE: a web tool delineating Mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data. *J Clin Microbiol* 53(6):1908–1914
19. Hunt M, Bradley P, Lapiere SG, Heys S, Thomsit M, Hall MB et al (2019) Antibiotic resistance prediction for Mycobacterium tuberculosis from genome sequence data with Mykrobe [version 1; peer review: 2 approved, 1 approved with reservations]. *Wellcome Open Res* 4
20. Sekizuka T, Yamashita A, Murase Y, Iwamoto T, Mitarai S, Kato S et al (2015) TGS-TB: total genotyping solution for Mycobacterium tuberculosis using short-read whole-genome sequencing. *PLoS One* 10(11):1–12
21. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S (2014) KvarQ: Targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics* 15(1):1–12
22. Taki T, Seki K, Wakabayashi Y, Morishige Y, Sekizuka T, Yamashita A et al (2019) Whole-genome sequencing-based epidemiological analysis of anti-tuberculosis drug resistance genes in Japan in 2007: application of the Genome Research for Asian Tuberculosis (GrEAT) database. *Sci Rep* 9(1):1–8
23. Iwai H, Kato-Miyazawa M, Kirikae T, Miyoshi-Akiyama T (2015) CASTB (the comprehensive analysis server for the Mycobacterium tuberculosis complex): a publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates. *Tuberculosis (Edinb)* [Internet] 95(6):843–844. Available from <https://doi.org/10.1016/j.tube.2015.09.002>
24. Dookie N, Rambaran S, Padayatchi N, Mahomed S, Naidoo K (2018) Evolution of drug resistance in Mycobacterium tuberculosis: a review on the molecular determinants of resistance and implications for personalized care. *J Antimicrob Chemother* 73(5):1138–1151

25. Macalino SJY, Billones JB, Organo VG, Carrillo MCO (2020) In silico strategies in tuberculosis drug discovery. *Molecules* 25(3):1–32
26. Advani J, Verma R, Chatterjee O, Pachouri PK, Upadhyay P, Singh R et al (2019) Whole genome sequencing of Mycobacterium tuberculosis clinical isolates from India reveals genetic heterogeneity and region-specific variations that might affect drug susceptibility. *Front Microbiol* 10(Feb):1–15
27. Veziris N, Bernard C, Guglielmetti L, Le Du D, Marigot-Outtandy D, Jaspard M et al (2017) Rapid emergence of Mycobacterium tuberculosis bedaquiline resistance: lessons to avoid repeating past errors. *Eur Respir J* [Internet] 49(3):1601719. Available from <http://erj.ersjournals.com/lookup/doi/10.1183/13993003.01719-2016>
28. Andries K, Villellas C, Coeck N, Thys K, Gevers T, Vranckx L et al (2014) Acquired resistance of Mycobacterium tuberculosis to bedaquiline. *PLoS One* 9(7):1–11
29. Milano A, Pasca MR, Proveddi R, Lucarelli AP, Manina G, Luisa de Jesus Lopes Ribeiro A et al (2009) Azole resistance in Mycobacterium tuberculosis is mediated by the MmpS5-MmpL5 efflux system. *Tuberculosis* [Internet] 89(1):84–90. Available from <https://doi.org/10.1016/j.tube.2008.08.003>
30. Ioerger TR, O'Malley T, Liao R, Guinn KM, Hickey MJ, Mohaideen N et al (2013) Identification of new drug targets and resistance mechanisms in Mycobacterium tuberculosis. Kremer L (ed). *PLoS One* [Internet] 8(9):e75245. Available from <https://dx.plos.org/10.1371/journal.pone.0075245>
31. Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, Barry CE et al (2017) Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nat Genet* [Internet] 49(3):395–402. Available from <http://www.nature.com/articles/ng.3767>
32. Hirani N, Joshi A, Anand S, Chowdhary A, Ganesan K, Agarwal M et al (2020) Detection of a novel mutation in the rpoB gene in a multidrug resistant Mycobacterium tuberculosis isolate using whole genome next generation sequencing. *J Glob Antimicrob Resist* [Internet] 22:270–274. Available from <https://doi.org/10.1016/j.jgar.2020.03.004>
33. Zhu L, Zhong J, Jia X, Liu G, Kang Y, Dong M et al (2016) Precision methylome characterization of Mycobacterium tuberculosis complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. *Nucleic Acids Res* 44(2):730–743
34. Shell SS, Prestwich EG, Baek SH, Shah RR, Sasseti CM, Dedon PC et al (2013) DNA methylation impacts gene expression and ensures hypoxic survival of Mycobacterium tuberculosis. *PLoS Pathog* 9(7):24–28
35. Seo S, Renaud C, Kuypers JM, Chiu CY, Huang ML, Samayoa E et al (2015) Idiopathic pneumonia syndrome after hematopoietic cell transplantation: evidence of occult infectious etiologies. *Blood* 125(24):3789–3797
36. Li Y, Sun B, Tang X, Liu Y-L, He H-Y, Li X-Y et al (2020) Application of metagenomic next-generation sequencing for bronchoalveolar lavage diagnostics in critically ill patients. *Eur J Clin Microbiol Infect Dis* 39(2):369–374
37. Zhou X, Wu H, Ruan Q, Jiang N, Chen X, Shen Y et al (2019) Clinical evaluation of diagnosis efficacy of active Mycobacterium tuberculosis complex infection via metagenomic next-generation sequencing of direct clinical samples. *Front Cell Infect Microbiol* 9(October):1–9
38. Trauner A, Liu Q, Via LE, Liu X, Ruan X, Liang L et al (2017) The within-host population dynamics of Mycobacterium tuberculosis vary with treatment efficacy. *Genome Biol* 18(1):1–17
39. Dheda K, Lenders L, Magombedze G, Srivastava S, Raj P, Arning E et al (2018) Drug-penetration gradients associated with acquired drug resistance in patients with tuberculosis. *Am J Respir Crit Care Med* 198(9):1208–1219
40. Bloembergen G V., Keller PM, Stucki D, Trauner A, Borrell S, Latshang T et al (2015) Acquired resistance to bedaquiline and delamanid in therapy for tuberculosis. *N Engl J Med* [Internet] 373(20):1986–1988. Available from <http://www.nejm.org/doi/10.1056/NEJMc1505196>

41. Eldholm V, Norheim G, von der Lippe B, Kinander W, Dahle UR, Caugant DA et al (2014) Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible ancestor in a single patient. *Genome Biol* [Internet] 15(11):490. Available from <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0490-3>
42. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ et al (2013) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis* [Internet] 13(2):137–146. Available from [https://doi.org/10.1016/S1473-3099\(12\)70277-3](https://doi.org/10.1016/S1473-3099(12)70277-3)
43. Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniewski F (2016) Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in London: a retrospective observational study. *PLoS Med* 13(10):1–18
44. Witney AA, Bateson ALE, Jindani A, Phillips PPJ, Coleman D, Stoker NG et al (2017) Use of whole-genome sequencing to distinguish relapse from reinfection in a completed tuberculosis clinical trial. *BMC Med* 15(1):1–13
45. Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y et al (2012) Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. *J Infect Dis* 206(11):1724–1733
46. Leung KSS, Siu GKH, Tam KKG, To SWC, Rajwani R, Ho PL et al (2017) Comparative genomic analysis of two clonally related multidrug resistant *Mycobacterium tuberculosis* by single molecule real time sequencing. *Front Cell Infect Microbiol* 7(Nov)
47. Cabibbe AM, Spitaleri A, Battaglia S, Colman RE, Suresh A, Uplekar S et al (2020) Application of targeted Next Generation Sequencing assay on a portable sequencing platform for culture-free detection of drug resistant tuberculosis from clinical samples. *J Clin Microbiol*. <https://doi.org/10.1128/JCM.00632-20>
48. Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J et al (2018) Direct whole-genome sequencing of sputum accurately identifies drug-resistant *Mycobacterium tuberculosis* faster than MGIT culture sequencing. *J Clin Microbiol* 56(8):1–11
49. Ko DH, Lee EJ, Lee SK, Kim HS, Shin SY, Hyun J et al (2019) Application of next-generation sequencing to detect variants of drug-resistant *Mycobacterium tuberculosis*: genotype-phenotype correlation. *Ann Clin Microbiol Antimicrob* [Internet] 18(1):1–8. Available from <https://doi.org/10.1186/s12941-018-0300-y>
50. Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, Meehan CJ et al (2019) GWAS for quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals resistance genes and regulatory regions. *Nat Commun* [Internet] 10(1). Available from <https://doi.org/10.1038/s41467-019-10110-6>
51. Oppong YEA, Phelan J, Perdigão J, MacHado D, Miranda A, Portugal I et al (2019) Genome-wide analysis of *Mycobacterium tuberculosis* polymorphisms reveals lineage-specific associations with drug resistance. *BMC Genomics* 20(1):1–15
52. Meehan CJ, Moris P, Kohl TA, Pečerska J, Akter S, Merker M et al (2018) The relationship between transmission time and clustering methods in *Mycobacterium tuberculosis* epidemiology. *EBioMedicine* 37(2018):410–416
53. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodtkin E et al (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 364(8):730–739
54. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J et al (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med* 10(2)
55. Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V et al (2015) Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. *Nat Commun* 6(May):7119
56. Zumla AI, Gillespie SH, Hoelscher M, Philips PPJ, Cole ST, Abubakar I et al (2014) New antituberculosis drugs, regimens, and adjunct therapies: needs, advances, and future prospects.

- Lancet Infect Dis [Internet] 14(4):327–340. Available from [https://doi.org/10.1016/S1473-3099\(13\)70328-1](https://doi.org/10.1016/S1473-3099(13)70328-1)
57. Hoffmann H, Kohl TA, Hofmann-Thiel S, Merker M, Beckert P, Jaton K et al (2016) Delamanid and bedaquiline resistance in mycobacterium tuberculosis ancestral Beijing genotype causing extensively drug-resistant tuberculosis in a Tibetan refugee. *Am J Respir Crit Care Med* 193(3):337–340
 58. Ramirez LMN, Vargas KQ, Diaz G (2020) Whole genome sequencing for the analysis of drug resistant strains of mycobacterium tuberculosis: a systematic review for bedaquiline and delamanid. *Antibiotics* 9(3)
 59. Iketleng T, Lessells R, Dlamini MT, Mogashoa T, Mupfumi L, Moyo S et al (2018) Mycobacterium tuberculosis next-generation whole genome sequencing: opportunities and challenges. *Tuberc Res Treat* 2018:1–8
 60. Dlamini MT, Lessells R, Iketleng T, de Oliveira T (2019) Whole genome sequencing for drug-resistant tuberculosis management in South Africa: what gaps would this address and what are the challenges to implementation? *J Clin Tuberc Other Mycobact Dis* [Internet] 16:100115. Available from <https://doi.org/10.1016/j.jctube.2019.100115>
 61. Votintseva AA, Pankhurst LJ, Anson LW, Morgan MR, Gascoyne-Binzi D, Walker TM et al (2015) Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. *J Clin Microbiol* 53(4):1137–1143
 62. Buermans HPJ, den Dunnen JT (2014) Next generation sequencing technology: advances and applications. *Biochim Biophys Acta Mol Basis Dis* [Internet] 1842(10):1932–1941. Available from <https://doi.org/10.1016/j.bbadis.2014.06.015>
 63. Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA et al (2013) Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. *N Engl J Med* [Internet] 369(3):290–292. Available from <http://www.nejm.org/doi/10.1056/NEJMc1215305>
 64. Kato-Maeda M, Ho C, Passarelli B, Banaei N, Grinsdale J, Flores L et al (2013) Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. *PLoS One* 8(3):1–8
 65. WHO (2018) The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in *Mycobacterium tuberculosis* complex: technical guide, vol 54
 66. Levy SE, Myers RM (2016) Advancements in next-generation sequencing. *Annu Rev Genomics Hum Genet* 17:95–115
 67. Besser J, Carleton HA, Gerner-Smith P, Lindsey RL, Trees E (2018) Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect* [Internet] 24(4):335–341. Available from <https://doi.org/10.1016/j.cmi.2017.10.013>
 68. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR et al (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* [Internet] 13(341):341. Available from <http://www.ncbi.nlm.nih.gov/pubmed/22827831>
 69. Dohál M, Porvazník I, Pršo K, Rasmussen EM, Solovič I, Mokry J (2020) Whole-genome sequencing and *Mycobacterium tuberculosis*: challenges in sample preparation and sequencing data analysis. *Tuberculosis (Edinb)* [Internet] 123:101946. Available from <https://doi.org/10.1016/j.tube.2020.101946>
 70. Cabibbe AM, Walker TM, Niemann S, Cirillo DM (2018) Whole genome sequencing of *Mycobacterium tuberculosis*. *Eur Respir J* [Internet] 52(5):1–5. Available from <https://doi.org/10.1183/13993003.01163-2018>
 71. Bainomugisa A, Duarte T, Lavu E, Pandey S, Coulter C, Marais BJ et al (2018) A complete high-quality MinION nanopore assembly of an extensively drug-resistant *Mycobacterium tuberculosis* Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. *Microb Genomics* [Internet] 4(7). Available from <https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000188>

72. Menardo F, Duchêne S, Brites D, Gagneux S (2019) The molecular clock of mycobacterium tuberculosis. *PLoS Pathog* 15(9):1–24
73. Kühnert D, Coscolla M, Brites D, Stucki D, Metcalfe J, Fenner L et al (2018) Tuberculosis outbreak investigation using phylodynamic analysis. *Epidemics* [Internet] 25:47–53. Available from <https://pubmed.ncbi.nlm.nih.gov/29880306>
74. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O et al (2020) A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the African Great Lakes region. *Nat Commun* [Internet] 11(1):1–11. Available from <https://doi.org/10.1038/s41467-020-16626-6>
75. Séraphin MN, Norman A, Rasmussen EM, Gerace AM, Chiribau CB, Rowlinson MC et al (2019) Direct transmission of within-host *Mycobacterium tuberculosis* diversity to secondary cases can lead to variable between-host heterogeneity without de novo mutation: a genomic investigation. *EBioMedicine* 47:293–300



Deciphering the Role of Epigenetic Reprogramming in Host-Pathogen Interactions

3

Amandeep Kaur Kang, Andrew M. Lynn, and Uma Dhawan

Abstract

The success of a pathogen within the host depends on various extrinsic factors that work in a synergistic mechanism to promote pathogenesis. One such factor is driven by the changes observed within the host genome, providing survival and establishment of pathogens inside the host. Pathogens are also known for establishing their intracellular niche within the host by mimicking the host enzymes and immune system for survival. Understanding the strategies used by pathogens to intervene in host genetic machinery for pathogenesis is important for creating successful targets and personalized drugs to counterbalance their effects. Accumulation of omics data and simultaneous development of bioinformatics analysis tools have allowed researchers to understand the interplay between prokaryotic and eukaryotic cells through the multi-omics approach. This permits a better understanding of diseases associated with host-parasite interactions and subsequent development of personalized medicines as therapeutics.

Keywords

Epigenetic modification · DNA methylation · Host signaling pathways · Pathogenic plasticity · Omics technologies · Next generation sequencing · Third generation sequencing

A. K. Kang · U. Dhawan (✉)

Department of Biomedical Science, Bhaskaracharya College of Applied Sciences, University of Delhi, New Delhi, India

e-mail: uma.dhawan@bcas.du.ac.in

A. M. Lynn

School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_3

3.1 Introduction

Pathoepigenetics is an emerging field of microbiology which deals with the epigenetic changes involved in host-pathogen interactions that are vital for the survival and multiplication of pathogens to induce infection within the host. More than 1400 species of human pathogens including viruses, bacteria, protozoans, and helminths have been observed. In order to thrive, they have been evolving along with humans, evading the innate and adaptive immune responses, thereby conquering their host. Understanding the molecular mechanisms involved in epigenetic changes triggered by pathogens is important to demonstrate the signaling pathways affected during infection. In order to beat the devastating infectious diseases, humans have been coevolving with pathogens by altering their genome to co-adapt. The most significant evolutionary machinery consists of a major histocompatibility complex (MHC), which shows diversity within individuals and contains the memory of past infections. Innate and adaptive immune systems collaborate to counterbalance the effects of pathogens. In order to establish themselves, the infectious agents aim to attack the host's defense system. Several bacteria and viruses aim to alter the epigenetic machinery of the host. They have been shown to initiate reprogramming of the innate immune cells. The pathogenic effector molecules modulate histone and protein deacetylation to promote regulatory T cell (Treg) [1]. *Clostridium perfringens* and *Streptococcus pneumoniae* have been shown to secrete toxins, namely, perfringolysin and pneumolysin, respectively, which lead to phosphorylation of H3S10. *Listeria monocytogenes* have been shown to induce H3S10 phosphorylation and deacetylation of H3 and H4 histones, thereby altering the chromatin for pathogenesis. Other bacteria have been shown to spread their virulence by modulating HDAC1 family proteins which promote epigenetic tolerance against these microbes [2–7]. The potential role of microbial infections in allergic diseases and autoimmune diseases has also been linked to the modulation of epigenetic factors through altering mucosal surfaces and counterbalancing the innate defense system of the host [8]. Highlighting the potential virulence determinants that epigenetically modulate the host genome will provide an understanding for the development of therapeutics to evade the infection. The dynamic nature of environment-driven epigenetic plasticity has enabled the host and pathogen to find new strategies for the survival of the fittest.

3.1.1 The Epigenetic Code

While the human genome sequence has transformed our understanding of human biology, it is not just the sequence of our DNA that matters, but how we use it and how are things executed within a cellular machinery. Why are some genes activated in certain cell types while others are silenced? Which factors work in synergy to regulate these differentially expressed genes? What properties differentiate a nerve cell from a smooth muscle cell? The key to this is epigenetics. Epigenetic changes are heritable through cell divisions and reversible and hold the potential to be

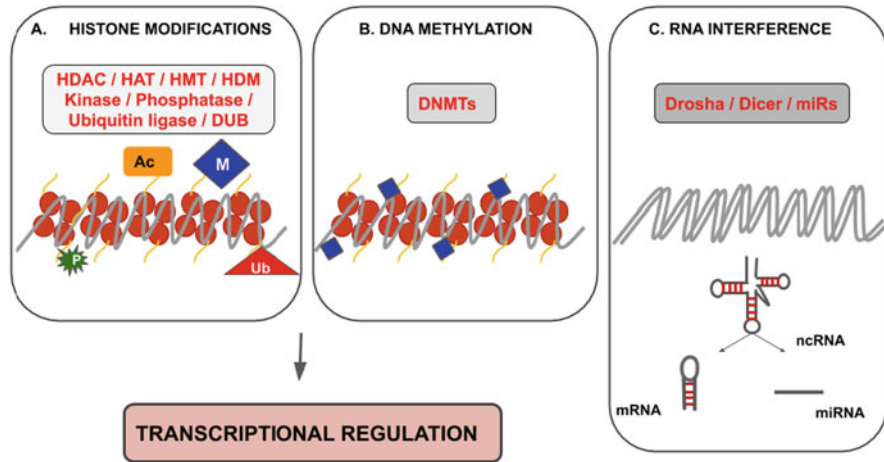


Fig. 3.1 Epigenetic modifications are illustrated here which lead to chromatin remodeling into active or inactive states. (a) DNA is wrapped around nucleosomes which are made of histone proteins which are prone to epigenetic alterations. Histone modifications include acetylation/deacetylation by HAT and HDAC, histone methylation/demethylation by HMT, and HDM and histone phosphorylation/dephosphorylation by kinases and phosphatases, respectively, and ubiquitination by ubiquitin ligase which adds ubiquitin to histones and deubiquitination by DUBs. (b) DNA methylation includes the addition of methyl groups by DNMTs which leads to transcriptional repression or silencing. (c) Epigenetic modifications through RNA interference by cleavage of ncRNAs into mRNAs and miRNAs. These miRNAs sit on the 3' end of UTRs of mRNAs and thus prevent translation. Ac, acetyl; DUB, deubiquitinases; HDAC, histone deacetylases; HAT, histone acetyltransferases; HMT, histone methyltransferases; HDM, histone demethylases; DNMTs, DNA methyltransferases; M, methyl; miRNA, microRNA; mRNA, messenger RNA; ncRNAs, noncoding RNAs; P, phosphate; Ub, ubiquitin

manipulated therapeutically. These modifications are sensitive to the environment. Epigenetics is the study of factors associated with behavioral and environmentally induced heritable changes within the gene expression that arise from chemical modifications of DNA or histone proteins. These changes are known to alter the phenotype of an organism without changing the genotype.

Molecular analysis shows that epigenetic changes comprise covalent modifications like DNA and histone methylation, phosphorylation, ubiquitination, SUMOylation, ADP ribosylation, citrullination, and acetylation [9, 10]. Eukaryotic DNA is tightly wrapped around the histone proteins. Majorly studied eukaryotic epigenetic mechanisms comprise methylation of cytosine residues in DNA and histone modifications that regulate nucleosome stability. Posttranslational modifications (PMTs) like histone methylation/demethylation and acetylation/deacetylation result in changes associated with the switching on and off of genes. These chromatin modifications are modulated by enzymes known as “writers,” like certain kinases, histone acetyltransferases (HATs), and histone methyltransferases (HMTs), and “erasers” like phosphatases, histone deacetylases (HDACs), and histone demethylases (HDMs) [11] (Fig. 3.1). Bacteria, on the other hand, lacks

histones; thus, the major epigenetic modifications include adenine and cytosine methylation which regulates gene expression and consists of a restriction-modification system which protects bacterial DNA from cleavage [12].

3.1.2 Epigenetic Reprogramming Driven by Extrinsic Factors

Once thought to be simply heritable, epigenetic changes are those extrinsic changes which are now considered to modulate the intrinsic environment throughout the organism's lifespan during cellular differentiation. These extrinsic changes include physical environmental stresses, lifestyle, nutritional factors, environmental toxins, and pharmacological treatments an organism undergoes during its lifetime. The prevailing environmental conditions can modulate the genetic expression of a trait through epigenetic alterations providing plasticity to the organism for adapting to the environment [13]. Thus, epigenetic changes ensure the induction of alternative phenotypes without an actual change in the genotype of the organism. Understanding the interactions between these environmental factors and their impact on the epigenome can help us predict the healthy or disease-associated phenotypes of the organism [14]. The environment-induced epigenetic changes are also dependent on the titer of infection or bacterial load and the duration of infection [15]. For this, the bacteria must establish itself in the microenvironment of the host by evading the host defense mechanisms. The higher the bacterial load and duration of infection within the host, the greater will be the epigenetic changes.

3.1.2.1 DNA Methylation

It is an epigenetic change marked by the addition of a methyl group to bases in the DNA sequence. The most frequently studied methylation is of the C5 position on cytosine bases using DNA methyltransferases (DNMTs) as writers [16, 17]. CpG methylation is the most dominant form of methylation in eukaryotes which can suppress transcription by blocking DNA binding by transcription factors, while in bacteria, methylation of the adenine residues is the main epigenetic signal. Immunoprecipitation and bisulfite-based techniques can be used together with microarrays or next-generation sequencing to decipher the genomic regions that are epigenetically modified. Recently, changes in DNA methylation induced by *E. coli* were observed in porcine cells where DNA methylation was shown to be majorly affected in immune response genes [18]. *Helicobacter pylori* infection can cause DNA methylation in the human gastric mucosa within genes associated with gastric cancer [19–21]. Within the uroepithelial cells, *E. coli* infection results in the upregulation of DNMT expression which induces CpG methylation which enables pathogen persistence within the host [22].

3.1.2.2 Histone Modification

Posttranslational histone remodeling can be achieved in different ways like histone acetylation, methylation, phosphorylation, and ubiquitination. Acetylation is catalyzed by histone acetyltransferase enzymes (HATs) which add an acetyl group

to the positively charged lysine amino acids within the histone tails, thus masking the positive charge. Transcriptionally permissive modifications include H3/H4 histone acetylation at the ϵ -amino group of lysine residues [23]. In contrast, deacetylation of histones is carried out by HDACs (histone deacetylase enzymes) and correlates with CpG methylation and inactive state of chromatin, thereby repressing transcription. HDACs are also regulated by phosphorylation, acetylation, and SUMOylation. Histone methylation includes modifications like H3K9me and H3K27me which can be related to chromatin repression [24], whereas H3K4me3, H3K36me3, H3S10p, and H3K14ac modifications are related with chromatin activation [25]. Furthermore, methylation can occur on ϵ -amino groups of arginine or lysine amino acids catalyzed by histone methyltransferases, but without any change in the charge of amino acids. This modification can be associated with both active and repressive gene transcription [26]. RV1988, a methyltransferase secreted by *M. tuberculosis*, methylates histone H3 at residue R42, promoting gene activation [27]. Mass spectrometry and genomics-based techniques such as ChIP-seq and ChIP-chip can be applied to detect specific regions of the genome associated with histone modifications. Bacterial histone acetylation/deacetylation and phosphorylation/dephosphorylation are involved in the alteration of microbe-associated molecular patterns and virulence factors involved in host-bacteria interactions. Histone methylation is the major histone modification targeted by bacteria [28]. SET domain proteins from various bacteria, like *Burkholderia thailandensis* and *Bacillus anthracis*, have been shown to cause histone methylation for transcriptional modification in the host [29].

3.1.2.3 RNA-Based Silencing

Gene regulation can also be achieved by antisense transcripts, by noncoding RNAs, or through RNA interference. RNA-based silencing alters the gene expression by triggering histone modifications or DNA methylation resulting in heterochromatin formation [25]. Within the nucleus, different long noncoding RNAs regulate the epigenetic status of various protein-coding genes, modifying gene transcription by recruiting chromatin remodeling complexes [30]. Long ncRNAs and sRNAs have been reported to participate in various regulatory processes involving chromatin or transcriptional regulation, nuclear architecture, and RNA processing [31–33]. lncRNAs alter the epigenetic processes by remodeling chromatin structure, while miRNAs are known to regulate DNMT expression in somatic cells and during embryonic development [34]. IsrM, one of the sRNAs of *Salmonella*, has been shown to promote bacterial invasion in hosts [35]. Recently, Gao et al. identified the survival strategy of bacteria *Edwardsiella tarda* within the intestine of humans by modulating sRNAs for establishment in hosts [36].

3.1.3 The Epigenetic Bridge of Survival: How Pathogens Change the Epigenetic Signals to Modulate Gene Transcription and Translation

The epigenetics of host-pathogen interactions aims to understand the dynamic and plastic nature of pathogenicity which directly links to the successful alteration of the host environment for survival and transmission of pathogens. Pathogens conquer the epigenetic signaling by altering the epigenetic modifications of genes associated with virulence processes, which allows their colonization, replication, and dissemination within the host. Bacteria secrete effector molecules like nucleomodulins [28] which enter the host nucleus and hijack the epigenetic machinery by manipulating the epigenetic factors, sRNAs, ncRNAs, and mRNAs [37] (Fig. 3.2).

3.1.3.1 The Bacterial Epigenome

Bacteria also utilize epigenetic modifications for various cellular functions like DNA replication, DNA repair, bacteriophage packaging, transposition, chromosomal segregation, transcriptional regulation, and interestingly, alteration of host cellular environment for pathogenicity. Adenine methylation is one of the extensively studied epigenetic modifications in bacteria which is reported to be regulated by DNA adenine methyltransferase (Dam) in *E. coli* and *Gammaproteobacteria* [38] while cell cycle-regulated methyltransferase (CcrM) has been studied in *Alphaproteobacteria* [39, 40]. DNA adenine methylation was found to be vital for *Salmonella* species [41, 42]. Restriction-modification systems in bacteria are known to have their own DNA methyltransferases which protect self-DNA from degradation after cleavage by restriction-modification enzymes [43]. Bacteria are shown to undergo a tremendous amount of phase variation which involves random and reversible switching of gene expression resulting in a wide variety of phenotypic cell variants [44, 45] known as phasevariations [46]. These phasevariations exhibit a heterogeneous mixed expression state with the gene either in the “active” or “repressed” state. This equips bacteria for immune evasion by providing a better resistance strategy for colonization inside the host environment and escaping membrane-specific vaccines. Such bacteria are categorized under the human-adapted pathogens, most commonly studied in *E. coli*, *Haemophilus influenzae*, *Helicobacter pylori*, and *Salmonella* species [47, 48]. Such changes are mediated by methyltransferases of the restriction-modification system and Dam. An outer membrane protein antigen 43 (Ag43), encoded by the Agn43 gene in *Escherichia coli*, is important for biofilms and infection. It is controlled by phase variation mediated by two proteins, Dam and the oxidative stress regulator OxyR. The GATC sequences of the promoter region of Agn43 gene overlap with the OxyR binding site. The binding of OxyR to this regulatory region of Agn43 leads to transcription repression of Agn43. However, Dam methylation of GATC sequences results in the transcriptional activation of Agn43 by preventing OxyR binding [49]. Phase variation has been known to cause immune evasion in a wide variety of bacteria like *Streptococcus pneumoniae*, *Clostridioides difficile*, *Vibrio*, and *Haemophilus* [50–53].

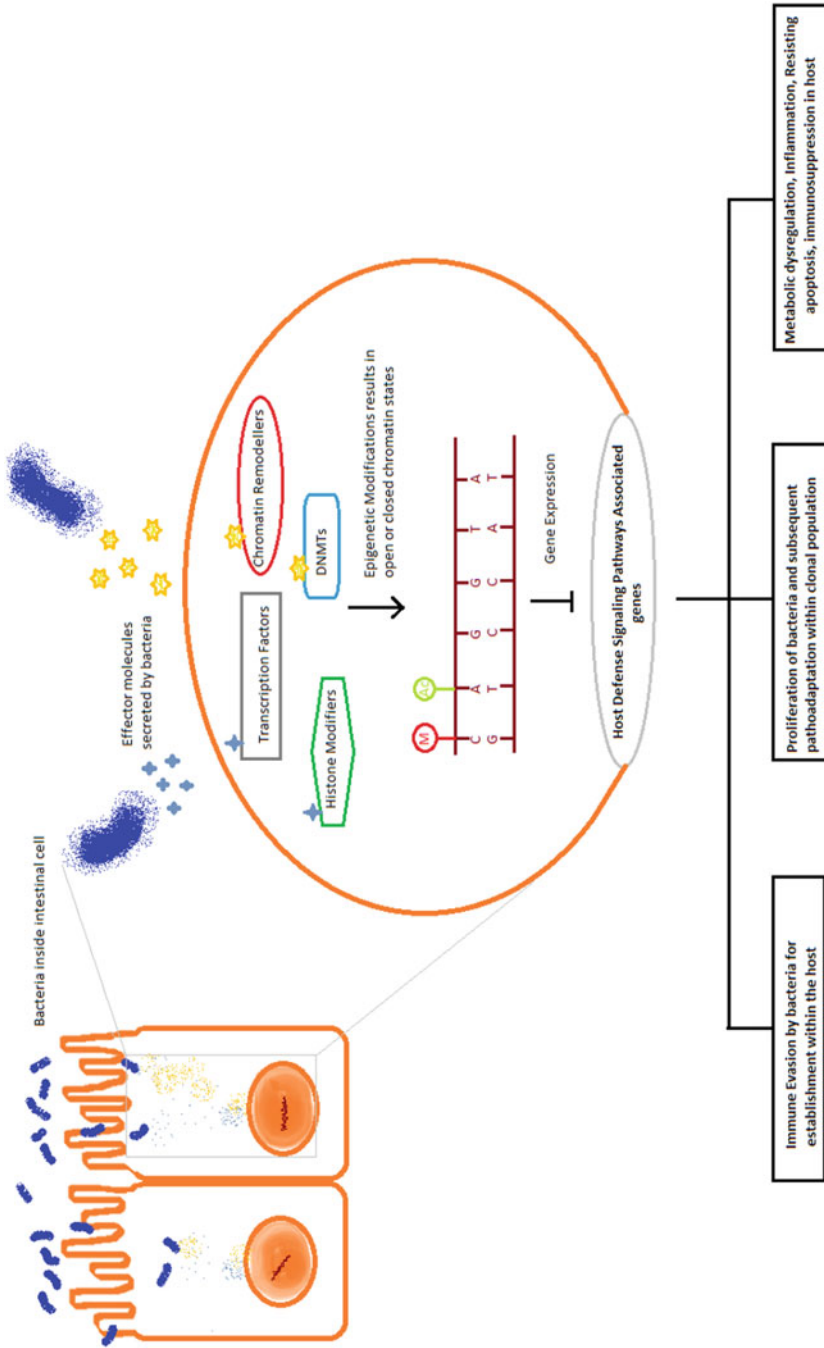


Fig. 3.2 The epigenetic modifications of bacteria within human intestinal cells. Bacteria secrete effector molecules like nucleomodulins which bind to transcription factors, chromatin remodelers, histone modifiers, DNMTs, etc. to regulate gene expression. Silencing or activation of genes associated with host defense pathways enable successful niche establishment, proliferation, adaptation, and survival of bacteria within the host

3.1.3.2 Pathogenic Plasticity

Bacterial genome plasticity contributes in shaping host-pathogen interactions for the colonization, invasion, survival, multiplication, and transmission of bacteria within the host. The challenges faced by pathogens within the diverse host environment elicit adaptive changes and mutations which can be observed morphologically and developmentally within the pathogens, thus rendering them protection from host defenses and therapeutic interventions. In order to facilitate survival in the host, bacteria acquired various strategies to terminate host cellular responses by altering host signaling pathways [54], targeting chromatin regulation, and modulating epigenetic marks. Bacteria encode certain effector molecules that modify host epigenetic machinery [55]. Protist *Plasmodium* has adapted to the host environment by exhibiting erythrocytic and hepatocytic stages which meet the pathogen's developmental requirements and enable it to survive longer within the host. These stages encode for genetic diversity and plasticity within the clonal population of pathogens. These patho-adaptive changes contribute to the fitness of pathogens. Similarly, bacteria undergo selective pressures within the host which allow genetic changes contributing to pathogenic plasticity. Within the same species, bacterial strains show variations in symptomatic and long-lasting asymptomatic cycles of infection. Pathogenic bacteria like *E. coli*, *M. tuberculosis*, and *S. typhi* can be asymptotically carried as a symbiont in hosts without showing any symptoms of infection [56]. They escape detection by hiding inside the macrophages within granulomas [57].

3.1.4 Host Signaling Pathways Altered for Pathogenicity

The effects of host-pathogen interaction revolve around alteration of host signaling cascades which are important for bacterial colonization in the host niche. For successful establishment, bacteria need to modify their defense system for evasion from the host, weaken the host immune system, and alter the host cellular machinery by mimicking host-like factors [54]. Certain bacteria have been shown to modify chromatin factors resulting in altered transcriptional regulation. In order to weaken the host defense system, bacteria aim to target the immune-specific signaling pathways. This works by altering the state of chromatin resulting in the conversion of euchromatin and heterochromatin or vice versa. Bacteria are involved in alteration of host MAPK, PI3K, and NF- κ B signaling cascades leading to downstream activation of kinases like AKT, IKK- α , and MSK which are involved in histone H3S10 phosphorylation and acetylation of H3K14 and H4K8 at the promoter regions of pro-inflammatory genes like IL-8 due to transcriptional repression [58, 59]. This results in the suppression of host inflammatory response against the invading bacteria [60, 61]. Gram-negative bacteria like *Shigella flexneri* have been shown to inhibit MAPK pathway and subsequent blocking of pro-inflammatory genes [62, 63]. The production of metabolites by bacteria leads to inhibition of chromatin-modifying enzymes in the host. One such metabolite, butyric acid, acts as an inhibitor of HDACs [64]. Also, certain bacteria like *Anaplasma*

phagocytophilum, *Ehrlichia*, and *Coxiella* have been shown to produce Ank-containing proteins which bind to the host nuclear chromatin. The motifs of bacterial Ank exhibit evolutionary homologies with eukaryotic counterparts. These result in altering protein-protein interaction and transcriptional regulation in the host imparting survival to the pathogen [65–67]. Differentially methylated CpGs in *E. coli*-infected porcine cells are composed of pro-inflammatory molecules like PAX5, AP4, IRF2, XBP1, and CREB with a significant reduction in DNA methyltransferases (DNMTs) which control the epigenetic modifications of the host [19].

3.2 Omics Technologies to Investigate Host-Pathogen Interactions

Traditional methods for diagnosing bacterial infections are composed of sensitive microbial cultures and isolation, followed by serological, immunological, and biochemical detection [68]. However, due to differences between in vivo host environments and in vitro cultures of bacteria, the host-pathogen interaction studies were incomplete. Also, detection of genetic, epigenetic, and metabolic differences initiated by pathogens was not possible through traditional culture and serological diagnosis [69, 70]. Molecular detection methods included real-time polymerase chain reaction, antimicrobial susceptibility testing, mass spectrometry (MS)-based methods [71–73], and immunoassays which are still considered gold standard methods for the identification of bacterial infections. However, due to insensitivity in the detection of certain species and strains, the diagnosis remains limited. These conventional diagnostic methods and molecular characterization methods have been successful in the identification of infections and controlling pandemics, but they are very laborious and time-consuming with poor resolution and specificity [74].

With the spread of infectious agents and increment in death rates as a result of bacterial infections, modernized technologies have gained popularity in high-throughput detection of these causative agents [75]. An advent of sequencing technologies have allowed researchers to understand the in vivo dynamics of pathogenesis [76]. With the revolution in high-throughput sequencing, whole genome sequencing has become a routine tool for clinical microbiology [77, 78]. The challenges provided by outbreaks of drug-resistant bacteria pose huge threats to the medical community. Therefore, it is important to understand the transmission, colonization, and establishment of pathogens within the host through genotypic tools. Due to greater diversity, strain-specific bacteria could not be identified through clinical diagnostic tests and first-generation sequencing methods. A more advanced second-generation sequencing platform permits bacterial genomes sequencing within hours. Whole genome sequencing and comparative genomics of *Escherichia coli* isolates showing diverse toxicity have been used to access the virulence of different strains. This data has been combined with epidemiological and phenotypic analysis to analyze the risk prediction during outbreaks. This was used to predict the marker genes for virulence of the pathogen using

GWAS studies [79, 80]. Sequencing technologies are rapidly improving. Third-generation sequencing platforms provide additional information with longer reads and accurate prediction of methylation sites within less time. This chapter mainly focuses on the methods used to predict epigenetic changes in bacterial infections (Table 3.1).

3.2.1 Epigenomic Techniques to Study Host-Pathogen Interactions

Technical challenges in studying the impact of bacterial load and associated changes in the intracellular environment of the host have been replaced with omics technologies. Over the last two decades, several assays have been designed for assessing the epigenetic changes. These are described in the following sections.

3.2.1.1 ChIP Assay

Chromatin immunoprecipitation assay monitors the epigenetic changes and transcriptional regulation associated with DNA-protein interactions [97]. ChIP assays use formaldehyde to crosslink DNA sequences and DNA-binding proteins in the form of complexes within the bacterial cells. This is followed by fragmentation of bacterial DNA and targeted immunoprecipitation of the resulting complexes. Being semiquantitative, ChIP assays have been used in combination with real-time polymerase chain reaction (ChIP-qPCR) to obtain a quantitative measurement of the amount of DNA of interest bound to protein. This can be validated with other transcriptional profiling methods like deep sequencing, qRT-PCR, and DNA microarrays for transcript-level studies. ChIP assays have been used to study gene regulation in the intracellular pathogens. Since intracellular bacteria have been known to regulate host gene expression by modifying chromatin and associated histone proteins, ChIP assays have been extensively used to study gut microbiota population in *Escherichia coli*, *Staphylococcus aureus*, and *Salmonella typhimurium* infections [98–102]. These gut-on-a-chip systems have been used to understand the symbiotic associations between the human gut and microbiota [103]. ChIP microarray was used in combination with luciferase reporter assay for studying the molecular basis of gastric tumorigenesis associated with *H. pylori* infection. Methylation profiling identified hypermethylation in tumor suppressor FOXD3 promoter in mice and humans during *H. pylori* infections [104].

3.2.1.2 DNA Methylation Analysis

Traditional methods to identify DNA methylome used bisulfite treatment of DNA to determine methylation patterns in cells known as bisulfite sequencing (BS). This technique was considered a “gold standard” technology since it was extensively used to identify differentially methylated regions on CpG islands before the onset of NGS era, but it cannot be used to detect methylated adenine residues which are commonly altered in bacterial infections. Reduced representation bisulfite sequencing (RRBS) is a modification of bisulfite sequencing which combines BS with restriction enzymes to measure methylation levels on CpG sequences. RRBS in combination

Table 3.1 Omics technologies to study host-pathogen interactions

Type of approach	Description and use	Year	References
<i>Serological diagnosis</i>			
Flocculation tests	Involves flocculation or precipitation of antigen-antibody interactions	1876	[81]
Enzyme-linked immunosorbent assays (ELISAs)	Used to detect the presence or absence of microbial antigens using fluorescent or chemiluminescent or colorimetric signal readouts and quantify the signal	1971	[82]
Chemiluminescence immunoassays	Used to detect light signals which are emitted through the chemical reaction between probes or enzymes that are bound to specific antibodies	1995	[83]
MS proteomics	Involves isolation of bacterial pathogens from host cells followed by enzymatic digestion of proteins and resulting peptides are used for quantification and analysis with mass-spectrometry	1898	[84]
<i>Chromatin immunoprecipitation studies</i>			
N-ChIP	Uses unfixed native chromatin which is digested by nuclease yielding efficient immunoprecipitation of DNA. It is used to study tightly bound histone proteins	2003	[85]
X-ChIP	Uses fixed chromatin which is fragmented by sonication and is mainly used to study nonhistone proteins	2000	[86]
ChIP Cloning	Based on cloning and sequencing of immunoprecipitated DNA obtained from the standard ChIP method	2002	[87]
ChIP-qPCR	ChIP is combined with qPCR to quantify the amount of DNA bound to protein	Early 2000s	[88]
ChIP-CpG microarray	It is used to target ChIP sequencing of CpG islands where transcription factors binding to promoters can be detected. This method uses a combination of ChIP-PCR and microarray to study histone modifications	2003	[89]
<i>DNA methylation analysis (NGS)</i>			
Bisulfite sequencing (BS)	Utilizes bisulfite treatment of DNA to decipher methylation patterns within the cells	1992	[90]
Reduced representation bisulfite sequencing (RRBS)	Combines BS with restriction enzymes to measure methylation levels on CpG sequences	2005	[91]
MeDIP sequencing	Uses antibodies for the enrichment of differentially methylated regions	2005	[92]
Oxidative BS or oxBS-Seq	A modification of BS which can differentiate between 5-methylcytosine and 5-hydroxymethylcytosine after oxidizing DNA to form 5-formylcytosine	2012	[93]

(continued)

Table 3.1 (continued)

Type of approach	Description and use	Year	References
TAB-seq	Modification of BS which glucosylates 5-hydroxymethylcytosine and utilizes TET enzymes to convert 5-methylcytosine to 5-formylcytosine	2012	[94]
<i>Third generation sequencing</i>			
Single-molecule real-time (SMRT) DNA sequencing	To identify altered methyltransferases in bacterial infections along with positions of DNA modifications and has been successfully used in sequencing bacterial methylomes	2009	[95]
Nanopore MinION sequencing	It is used to identify methylated adenine and cytosine residues in bacterial DNA	2014	[96]

with RNA-seq transcriptomic profiling has been used to identify the differentially methylated regions in *Mycobacterium bovis*-infected cattle where epigenetic changes as a result of infection created dysfunctional CD4(+) T lymphocytes which were unable to clear *Mycobacterium* infection [105]. MBD-seq or methylated-CpG binding protein and MeDIP sequencing or methylated DNA immunoprecipitation reaction utilize antibodies for the enrichment of differentially methylated regions with better sensitivity in low CpG dense regions. Integrated MeDIP-ChIP and transcriptome analysis have been used to identify novel methylated signatures in porcine *Escherichia coli* induced diarrhea where changes associated with DNA methylation were observed in immune responses related genes, thus suppressing the host immune system [106]. Whole genome bisulfite sequencing (WGBS) technologies were developed which provided genome coverage at a single-base resolution, but due to higher expenditure, it is not extensively used. Deep sequence coverage of low CpG dense regions was achieved at a cost-effective and more accurate method by methylation capture sequencing or MethylCap-Seq technology. Restriction enzyme-based methods like methyl-sensitive cut counting (MSCC) depend on the restriction enzyme (like Msp1) digestion of CCGG motifs. Other modifications of BS are oxidative BS or oxBS-Seq and TAB-seq which were developed in 2012 since the traditional BS methods could not differentiate between 5-methylcytosine and 5-hydroxymethylcytosine, a TET-mediated modification of methylated cytosine. Ox-BS libraries and TAB-seq or Tet-assisted bisulfite sequencing allow identification of differentially methylated and hydroxymethylated regions at a single-base resolution.

3.2.1.3 Third Generation Methylome Profiling Technologies

Current advances in sequencing technologies allow interpretation of individual DNA molecules and identification of associated base modifications. For an in-depth characterization of the bacterial methylome, the most common third generation platforms include single-molecule real-time (SMRT) DNA sequencing [107] and

Nanopore MinION [108–110] sequencing that allow direct readouts for DNA modifications at a single-base resolution.

Nanopore DNA Sequencing Technology

Nanopore DNA sequencing technology developed by Oxford Nanopore Technologies (ONT) exploits differences in ionic current that occurs when different nucleotide bases pass through genetically modified protein nanopores. Nanopore MinION has been used to characterize bacterial methylomes for the identification of methylated cytosine and adenine residues in the DNA [111]. De novo-based sequencing for Nanopore has not been done so far.

SMRT DNA Sequencing Technology

SMRT DNA sequencing technology was manufactured by Pacific Biosciences Inc. (PacBio), is able to identify altered methyltransferases in bacterial infections along with positions of DNA modifications, and has been successfully used in sequencing bacterial methylomes. The output of SMRT includes simultaneous generation of nucleotide sequence and bacterial DNA methylation signatures (5mC, 4mC, and 6mA) with the relatively high signal-to-noise ratio. SMRT was used to identify methylated adenine residues in *Escherichia coli*-infected cells [107]. SMRT technology has provided deeper insights in understanding phase-variable methyltransferases [112, 113] in various species of bacteria including *Helicobacter pylori* [46, 114], *Haemophilus influenzae* [115], *Neisseria meningitidis* [113], and *Campylobacter jejuni* [116].

3.2.1.4 Single-Cell Epigenomics

Investigating the role of single-cell epigenomics has gained popularity, and it is used for characterizing cellular identity, molecular function, and understanding the phenotypes which cannot be predicted solely by the genotype. Epigenetic alterations can be identified as early-stage biomarkers for understanding the pathogenicity of infection and its therapeutics. Most common single-cell methylome assays include reduced-representation bisulfite sequencing (scRRBS), single-cell whole genome Bisulfite sequencing (SC-WGBS), or single-cell bisulfite sequencing (scBS-seq) for the identification of DNA methylation patterns and single-cell chromatin immunoprecipitation sequencing (scChIP-seq) for transcription factor identification and histone modification detection; scDNase-seq and scATACseq have been used for understanding the chromatin state and scHiC for chromosome conformation capturing.

3.3 Conclusion

A systematic approach towards reduction of pathogenic load and prevention of risks associated with pathogens led to the development of the microbial risk assessment (MRA) tool. Assessing the microbiological load helps in estimating the public health risk by quantifying the extent of spread of a disease or transfer of pathogens

preventing epidemic-like situations. Characterization of the severity of an infectious disease by next generation omics can help in refining our knowledge of the virulence of the pathogen. NGS technologies and high-throughput data analysis have produced innovative technologies for interpreting and understanding complex healthcare attributes. These NGS technologies include RNA-seq, and the expansion of genomics, transcriptomics, metabolomics, and proteomics has enabled us to monitor the individual strategies used by the pathogens for establishment inside the host. Integrating multi-omics approaches with research data has helped us in understanding the host-pathogen interactions. Detection of factors, genes, mimicked enzymes, and signaling components causing the infection through comparative genomics and analysis of these factors as potential biomarkers for the disease can help in the quick prediction and personalized therapeutic development for each strain of pathogen.

3.4 Future Perspectives

Understanding how bacteria mediate multiple levels of cellular and molecular states is fundamental to biomedical research. Multi-omics data integration combines multiple datasets generated by diagnostic tools and sequencing platforms with statistical analysis and correlates this information with biological pathway databases in order to relate the molecular dynamics of a diseased phenotype. These strategies have been in progress with the advent of third-generation sequencing technologies and production of bioinformatics tools to enable high-throughput data generation and analysis. Numerous data repositories have been developed which include Roadmap Epigenomics, Ensembl, Omics Integrator, 3Omics, Panther, String, DAVID, GenExp, Epigenome Atlas, VANTED, ProMeTra, and IntegrOmics. High-dimensional omics data require sophisticated software tools for analysis. Pipelines for analyzing omics data have been advancing along with the data generation. For each dataset, there is dynamicity in the implementation of these pipelines with minor to major changes associated with parameter modifications. Dependency on bioinformatics tools and repositories poses new challenges for advancement in analyzing multi-omics data with a higher resolution. Third generation sequencing methods possess immense potential in uncovering the dynamics of host-pathogen interactions at the molecular, cellular, and tissue-specific level. Most of the DNA methylation aiming at understanding host-pathogen interactions investigated tissue samples. Due to limited biopsy samples, there is a need for noninvasive DNA methylation methods for the detection of epigenetic modifications. One such advancement is observed in single-cell epigenome sequencing technology which provides a basic picture of disease-associated changes in cellular populations infected with pathogens. If used in combination with single-cell transcriptome sequencing, single-cell epigenome sequencing will provide us a better understanding of the dynamics of host-pathogen interactions [117]. Researchers are now investigating cell-free DNA sequencing technologies which harbor body fluids like serum, urine, and plasma for sequencing [118]. Even though Nanopore technology and SMRT need additional improvements, they continue to be promising platforms

for the identification of novel methyltransferases and methylated sites. Third generation methylome studies in collaboration with transcriptome studies and microarray will produce thousands of highly accurate and novel isoforms which will enable us to understand the in vivo dynamics of host-pathogen interactions. Identification of stage-specific biomarkers will allow us to diagnose the infection at earlier stages. Integrating the biomarker information and multi-omics data as a systems biology approach will enable us to unravel the high complexity of the biological system with better delivery of personalized therapeutics or targeted interventional therapies.

Acknowledgment AKK would like to acknowledge the UGC, Government of India for SRF. UD is supported by the SERB-DST grant (ECR/2017/000605).

References

1. Denzer L, Schrotten H, Schwerk C (2020) From gene to protein-how bacterial virulence factors manipulate host gene expression during infection. *Int J Mol Sci* 21(10):3730. <https://doi.org/10.3390/ijms21103730>. Published 2020 May 25
2. Garcia BA (2009) Mass spectrometric analysis of histone variants and post-translational modifications. *Front Biosci* 1:142–153. PMID: 19482690
3. Yin L, Chung WO (2011) Epigenetic regulation of human β -defensin 2 and CC chemokine ligand 20 expression in gingival epithelial cells in response to oral bacteria. *Mucosal Immunol* 4(4):409–419. <https://doi.org/10.1038/mi.2010.83>. Epub 2011 Jan 19. PMID: 21248725; PMCID: PMC3118861
4. Rennoll-Bankert KE, Dumler JS (2012) Lessons from *Anaplasma phagocytophilum*: chromatin remodeling by bacterial effectors. *Infect Disord Drug Targets* 12(5):380–387. <https://doi.org/10.2174/187152612804142242>
5. Grabiec AM, Potempa J (2018) Epigenetic regulation in bacterial infections: targeting histone deacetylases. *Crit Rev Microbiol* 44(3):336–350. <https://doi.org/10.1080/1040841X.2017.1373063>. Epub 2017 Oct 3. PMID: 28971711; PMCID: PMC6109591
6. Bandyopadhyaya A, Tsurumi A, Maura D, Jeffrey KL, Rahme LG (2016) A quorum-sensing signal promotes host tolerance training through HDAC1-mediated epigenetic reprogramming. *Nat Microbiol* 1:16174. <https://doi.org/10.1038/nmicrobiol.2016.174>. Published 2016 Oct 3
7. Niller HH, Masa R, Venkei A, Mészáros S, Minarovits J (2017) Pathogenic mechanisms of intracellular bacteria. *Curr Opin Infect Dis* 30(3):309–315. <https://doi.org/10.1097/QCO.000000000000363>. PMID: 28134679
8. Niller HH, Wolf H, Minarovits J (2008) Regulation and dysregulation of Epstein-Barr virus latency: implications for the development of autoimmune diseases. *Autoimmunity* 41(4):298–328. <https://doi.org/10.1080/08916930802024772>. PMID: 18432410
9. Kouzarides T (2007) Chromatin modifications and their function. *Cell* 128(4):693–705. <https://doi.org/10.1016/j.cell.2007.02.005>. PMID: 17320507
10. Suganuma T, Workman JL (2011) Signals and combinatorial functions of histone modifications. *Annu Rev Biochem* 80:473–499. <https://doi.org/10.1146/annurev-biochem-061809-175347>. PMID: 21529160
11. Zhou Y, Kim J, Yuan X, Braun T (2011) Epigenetic modifications of stem cells: a paradigm for the control of cardiac progenitor cells. *Circ Res* 109(9):1067–1081. <https://doi.org/10.1161/CIRCRESAHA.111.243709>. PMID: 21998298
12. Willbanks A, Leary M, Greenshields M, Tyminski C, Heerboth S, Lapinska K, Haskins K, Sarkar S (2016) The evolution of epigenetics: from prokaryotes to humans and its biological consequences. *Genet Epigenet* 8:25–36. <https://doi.org/10.4137/GEG.S31863>. PMID: 27512339; PMCID: PMC4973776

13. Gluckman PD, Hanson MA, Beedle AS (2007) Non-genomic transgenerational inheritance of disease risk. *Bioessays* 29(2):145–154. <https://doi.org/10.1002/bies.20522>. PMID: 17226802
14. Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* 33(Suppl):245–254. <https://doi.org/10.1038/ng1089>. PMID: 12610534
15. Ho SM, Johnson A, Tarapore P, Janakiram V, Zhang X, Leung YK (2012) Environmental epigenetics and its implication on disease risk and health outcomes [published correction appears in *ILAR J* 2017 Dec 15;58(3):413]. *ILAR J* 53(3–4):289–305. <https://doi.org/10.1093/ilar.53.3-4.289>
16. Klose RJ, Bird AP (2006) Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci* 31(2):89–97. <https://doi.org/10.1016/j.tibs.2005.12.008>. Epub 2006 Jan 5. PMID: 16403636
17. Chen ZX, Riggs AD (2011) DNA methylation and demethylation in mammals. *J Biol Chem* 286(21):18347–18353. <https://doi.org/10.1074/jbc.R110.205286>. Epub 2011 Mar 24. PMID: 21454628; PMCID: PMC3099650
18. Sajjanar B, Trakooljul N, Wimmers K, Ponsuksili S (2019) DNA methylation analysis of porcine mammary epithelial cells reveals differentially methylated loci associated with immune response against *Escherichia coli* challenge. *BMC Genomics* 20(1):623. <https://doi.org/10.1186/s12864-019-5976-7>. Published 2019 July 31
19. Maekita T, Nakazawa K, Mihara M, Nakajima T, Yanaoka K, Iguchi M, Arii K, Kaneda A, Tsukamoto T, Tatematsu M, Tamura G, Saito D, Sugimura T, Ichinose M, Ushijima T (2006) High levels of aberrant DNA methylation in *Helicobacter pylori*-infected gastric mucosae and its possible association with gastric cancer risk. *Clin Cancer Res* 12(3 Pt 1):989–995. <https://doi.org/10.1158/1078-0432.CCR-05-2096>. PMID: 16467114
20. Ding SZ, Goldberg JB, Hatakeyama M (2010) *Helicobacter pylori* infection, oncogenic pathways and epigenetic mechanisms in gastric carcinogenesis. *Future Oncol* 6(5):851–862. <https://doi.org/10.2217/fon.10.37>. PMID: 20465395; PMCID: PMC2882595
21. Ushijima T, Hattori N (2012) Molecular pathways: involvement of *Helicobacter pylori*-triggered inflammation in the formation of an epigenetic field defect, and its usefulness as cancer risk and exposure markers. *Clin Cancer Res* 18(4):923–929. <https://doi.org/10.1158/1078-0432.CCR-11-2011>. Epub 2011 Dec 28. PMID: 22205689
22. Tolg C, Sabha N, Cortese R, Panchal T, Ahsan A, Soliman A, Aitken KJ, Petronis A, Bägli DJ (2011) Uropathogenic *E. coli* infection provokes epigenetic downregulation of CDKN2A (p16INK4A) in uroepithelial cells. *Lab Invest* 91(6):825–836. <https://doi.org/10.1038/labinvest.2010.197>. Epub 2011 Jan 17. PMID: 21242958
23. Strahl BD, Allis CD (2000) The language of covalent histone modifications. *Nature* 403(6765):41–45. <https://doi.org/10.1038/47412>. PMID: 10638745
24. Bierne H, Hamon M, Cossart P (2012) Epigenetics and bacterial infections. *Cold Spring Harb Perspect Med* 2(12):a010272. <https://doi.org/10.1101/cshperspect.a010272>. Published 2012 Dec 1
25. Egger G, Liang G, Aparicio A, Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature* 429(6990):457–463. <https://doi.org/10.1038/nature02625>. PMID: 15164071
26. Hayakawa T, Nakayama J (2011) Physiological roles of class I HDAC complex and histone demethylase. *J Biomed Biotechnol* 2011:129383. <https://doi.org/10.1155/2011/129383>. Epub 2010 Oct 26. PMID: 21049000; PMCID: PMC2964911
27. Yaseen I, Kaur P, Nandicoori VK, Khosla S (2015) Mycobacteria modulate host epigenetic machinery by Rv1988 methylation of a non-tail arginine of histone H3. *Nat Commun* 6:8922. <https://doi.org/10.1038/ncomms9922>. PMID: 26568365
28. Dong W, Hamon MA (2020) Revealing eukaryotic histone-modifying mechanisms through bacterial infection. *Semin Immunopathol* 42(2):201–213. <https://doi.org/10.1007/s00281-019-00778-9>. Epub 2020 Feb 4. PMID: 32020336

29. Li T, Lu Q, Wang G, Xu H, Huang H, Cai T, Kan B, Ge J, Shao F (2013) SET-domain bacterial effectors target heterochromatin protein 1 to activate host rDNA transcription. *EMBO Rep* 14(8):733–740. <https://doi.org/10.1038/embor.2013.86>. Epub 2013 June 25. PMID: 23797873; PMCID: PMC3736128
30. Morlando M, Fatica A (2018) Alteration of epigenetic regulation by long noncoding RNAs in cancer. *Int J Mol Sci* 19(2):570. <https://doi.org/10.3390/ijms19020570>. Published 2018 Feb 14
31. Amaral PP, Dinger ME, Mattick JS (2013) Non-coding RNAs in homeostasis, disease and stress responses: an evolutionary perspective. *Brief Funct Genomics* 12(3):254–278. <https://doi.org/10.1093/bfgp/elt016>. PMID: 23709461
32. Quinn JJ, Chang HY (2016) Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* 17(1):47–62. <https://doi.org/10.1038/nrg.2015.10>. PMID: 26666209
33. Chen J, Wang Y, Wang C, Hu JF, Li W (2020) LncRNA functions as a new emerging epigenetic factor in determining the fate of stem cells. *Front Genet* 11:277. <https://doi.org/10.3389/fgene.2020.00277>. PMID: 32296461; PMCID: PMC7137347
34. Peschansky VJ, Wahlestedt C (2014) Non-coding RNAs as direct and indirect modulators of epigenetic regulation. *Epigenetics* 9(1):3–12. <https://doi.org/10.4161/epi.27473>. PMID: 24739571; PMCID: PMC3928183
35. Gong H, Vu GP, Bai Y, Chan E, Wu R, Yang E, Liu F, Lu S (2011) A Salmonella small non-coding RNA facilitates bacterial invasion and intracellular replication by modulating the expression of virulence factors. *PLoS Pathog* 7(9):e1002120. <https://doi.org/10.1371/journal.ppat.1002120>. Epub 2011 Sept 15. PMID: 21949647; PMCID: PMC3174252
36. Gao D, Zhang Y, Liu R, Fang Z, Lu C (2019) EsR240, a non-coding sRNA, is required for the resistance of *Edwardsiella tarda* to stresses in macrophages and for virulence. *Vet Microbiol* 231:254–263. <https://doi.org/10.1016/j.vetmic.2019.03.023>. Epub 2019 Mar 22. PMID: 30955819
37. Denzer L, Schrotten H, Schwerk C (2020) From gene to protein-how bacterial virulence factors manipulate host gene expression during infection. *Int J Mol Sci* 21(10):3730. <https://doi.org/10.3390/ijms21103730>. PMID: 32466312; PMCID: PMC7279228
38. Marinus MG (1996) Methylation of DNA in *Escherichia coli* and *Salmonella*. *Cell Mol Biol* 782–791
39. Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A et al (2016) The epigenomic landscape of prokaryotes. *PLoS Genet* 12:e1005854
40. Adhikari S, Curtis PD (2016) DNA methyltransferases and epigenetic regulation in bacteria. *FEMS Microbiol Rev* 40:575–591
41. Heithoff DM, Sinsheimer RL, Low DA, Mahan MJ (1999) An essential role for DNA adenine methylation in bacterial virulence. *Science* 284:967–970
42. Garcia-Del Portillo F, Pucciarelli MG, Casadesus J (1999) DNA adenine methylase mutants of *Salmonella typhimurium* show defects in protein secretion, cell invasion, and M cell cytotoxicity. *Proc Natl Acad Sci USA* 96:11578–11583
43. Bickle TA, Krüger DH (1993) Biology of DNA restriction. *Microbiol Rev* 57:434–450
44. Vasu K, Nagaraja V (2013) Diverse functions of restriction-modification systems in addition to cellular defence. *Microbiol Mol Biol Rev* 77:53–72
45. De Ste Croix M et al (2017) Phase-variable methylation and epigenetic regulation by type I restriction-modification systems. *FEMS Microbiol Rev* 41:S3–S15
46. Srikhanta YN, Gorrell RJ, Power PM, Tsyganov K, Boitano M, Clark TA, Korlach J, Hartland EL, Jennings MP, Kwok T (2017) Methylomic and phenotypic analysis of the ModH5 phase-variant of *Helicobacter pylori*. *Sci Rep* 7(1):16140. <https://doi.org/10.1038/s41598-017-15721-x>. PMID: 29170397; PMCID: PMC5700931
47. De Bolle X et al (2000) The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases. *Mol Microbiol* 35:211–222
48. de Vries N et al (2002) Transcriptional phase variation of a type III restriction-modification system in *Helicobacter pylori*. *J Bacteriol* 184:6615–6623

49. van der Woude MW, Henderson IR (2008) Regulation and function of Ag43 (flu). *Annu Rev Microbiol* 62:153–169. <https://doi.org/10.1146/annurev.micro.62.081307.162938>. PMID: 18785838
50. Li J, Zhang JR (2019) Phase variation of *Streptococcus pneumoniae*. *Microbiol Spectr* 7(1). <https://doi.org/10.1128/microbiolspec.GPP3-0005-2018>. PMID: 30737916
51. Anjuwon-Foster BR, Tamayo R (2018) Phase variation of *Clostridium difficile* virulence factors. *Gut Microbes* 9(1):76–83. <https://doi.org/10.1080/19490976.2017.1362526>. Epub 2017 Sept 21. Erratum for: Addendum to: Anjuwon-Foster BR, Tamayo R (2017) A genetic switch controls the production of flagella and toxins in *Clostridium difficile*. *PLoS Genet* 13(3):e1006701. PMID: 28806147; PMCID: PMC5914908
52. Hilton T, Rosche T, Froelich B, Smith B, Oliver J (2006) Capsular polysaccharide phase variation in *Vibrio vulnificus*. *Appl Environ Microbiol* 72(11):6986–6993. <https://doi.org/10.1128/AEM.00544-06>. Epub 2006 Aug 25. PMID: 16936057; PMCID: PMC1636181
53. Weiser JN, Love JM, Moxon ER (1989) The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. *Cell* 59(4):657–665. [https://doi.org/10.1016/0092-8674\(89\)90011-1](https://doi.org/10.1016/0092-8674(89)90011-1). PMID: 2479481
54. Brodsky IE, Medzhitov R (2009) Targeting of immune signalling networks by bacterial pathogens. *Nat Cell Biol* 11(5):521–526. <https://doi.org/10.1038/ncb0509-521>. PMID: 19404331
55. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, Koonin EV, Davis RW (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* 282(5389):754–759. <https://doi.org/10.1126/science.282.5389.754>. PMID: 9784136
56. Klemm P, Roos V, Ulett GC, Svanborg C, Schembri MA (2006) Molecular characterization of the *Escherichia coli* asymptomatic bacteriuria strain 83972: the taming of a pathogen. *Infect Immun* 74(1):781–785. <https://doi.org/10.1128/IAI.74.1.781-785.2006>. PMID: 16369040; PMCID: PMC1346676
57. Dobrindt U, Zdziarski J, Salvador E, Hacker J (2010) Bacterial genome plasticity and its impact on adaptation during persistent infection. *Int J Med Microbiol* 300(6):363–366. <https://doi.org/10.1016/j.ijmm.2010.04.010>. Epub 2010 May 7. PMID: 20452279
58. Schmeck B, Beermann W, van Laak V, Zahlten J, Opitz B, Witzenzath M, Hocke AC, Chakraborty T, Kracht M, Rosseau S, Suttorp N, Hippenstiel S (2005) Intracellular bacteria differentially regulated endothelial cytokine release by MAPK-dependent histone modification. *J Immunol* 175(5):2843–2850. <https://doi.org/10.4049/jimmunol.175.5.2843>. PMID: 16116170
59. Opitz B, Püschel A, Beermann W, Hocke AC, Förster S, Schmeck B, van Laak V, Chakraborty T, Suttorp N, Hippenstiel S (2006) *Listeria monocytogenes* activated p38 MAPK and induced IL-8 secretion in a nucleotide-binding oligomerization domain 1-dependent manner in endothelial cells. *J Immunol* 176(1):484–490. <https://doi.org/10.4049/jimmunol.176.1.484>. PMID: 16365441
60. Haller D, Holt L, Kim SC, Schwabe RF, Sartor RB, Jobin C (2003) Transforming growth factor-beta 1 inhibits non-pathogenic Gram negative bacteria-induced NF-kappa B recruitment to the interleukin-6 gene promoter in intestinal epithelial cells through modulation of histone acetylation. *J Biol Chem* 278(26):23851–23860. <https://doi.org/10.1074/jbc.M300075200>. Epub 2003 Apr 2. PMID: 12672795
61. Slevogt H, Schmeck B, Jonat C, Zahlten J, Beermann W, van Laak V, Opitz B, Dietel S, N'Guessan PD, Hippenstiel S, Suttorp N, Seybold J (2006) *Moraxella catarrhalis* induces inflammatory response of bronchial epithelial cells via MAPK and NF-kappaB activation and histone deacetylase activity reduction. *Am J Physiol Lung Cell Mol Physiol* 290(5):L818–L826. <https://doi.org/10.1152/ajplung.00428.2005>. Epub 2006 Jan 6. PMID: 16399788
62. Li H, Xu H, Zhou Y, Zhang J, Long C, Li S, Chen S, Zhou JM, Shao F (2007) The phosphothreonine lyase activity of a bacterial type III effector family. *Science* 315

- (5814):1000–1003. <https://doi.org/10.1126/science.1138960>. Erratum in: *Science* 2007 July 6;317(5834):43. PMID: 17303758
63. Brennan DF, Barford D (2009) Eliminylation: a post-translational modification catalyzed by phosphothreonine lyases. *Trends Biochem Sci* 34(3):108–114. <https://doi.org/10.1016/j.tibs.2008.11.005>. Epub 2009 Feb 21. PMID: 19233656
64. Riggs MG, Whittaker RG, Neumann JR, Ingram VM (1977) n-Butyrate causes histone modification in HeLa and Friend erythroleukaemia cells. *Nature* 268(5619):462–464. <https://doi.org/10.1038/268462a0>. PMID: 268489
65. Mosavi LK, Cammett TJ, Desrosiers DC, Peng ZY (2004) The ankyrin repeat as molecular architecture for protein recognition. *Protein Sci* 13(6):1435–1448. <https://doi.org/10.1110/ps.03554604>. PMID: 15152081; PMCID: PMC2279977
66. Park J, Kim KJ, Choi KS, Grab DJ, Dumler JS (2004) *Anaplasma phagocytophilum* Anka binds to granulocyte DNA and nuclear proteins. *Cell Microbiol* 6(8):743–751. <https://doi.org/10.1111/j.1462-5822.2004.00400.x>. PMID: 15236641
67. Zhu B, Nethery KA, Kuriakose JA, Wakeel A, Zhang X, McBride JW (2009) Nuclear translocated *Ehrlichia chaffeensis* ankyrin protein interacts with a specific adenine-rich motif of host promoter and intronic Alu elements. *Infect Immun* 77(10):4243–4255. <https://doi.org/10.1128/IAI.00376-09>. Epub 2009 Aug 3. PMID: 19651857; PMCID: PMC2747939
68. Rajapaksha P, Elbourne A, Gangadoo S, Brown R, Cozzolino D, Chapman J (2019) A review of methods for the detection of pathogenic microorganisms. *Analyst* 144:396–411
69. Schrader KN, Fernandez-Castro A, Cheung WK, Crandall CM, Abbott SL (2008) Evaluation of commercial antisera for *Salmonella* serotyping. *J Clin Microbiol* 46(2):685–688. <https://doi.org/10.1128/JCM.01808-07>. Epub 2007 Dec 19. PMID: 18094130; PMCID: PMC2238139
70. Prager R, Strutz U, Fruth A, Tschäpe H (2003) Subtyping of pathogenic *Escherichia coli* strains using flagellar (H)-antigens: serotyping versus *flhC* polymorphisms. *Int J Med Microbiol* 292(7-8):477–486. <https://doi.org/10.1078/1438-4221-00226>. PMID: 12635930
71. Meyer C, Stolle A, Fredriksson-Ahomaa M (2011) Comparison of broth microdilution and disk diffusion test for antimicrobial resistance testing in *Yersinia enterocolitica* 4/O:3 strains. *Microb Drug Resist* 17(3):479–484. <https://doi.org/10.1089/mdr.2011.0012>. Epub 2011 May 13. PMID: 21568753
72. Lee M, Chung HS (2015) Different antimicrobial susceptibility testing methods to detect ertapenem resistance in Enterobacteriaceae: VITEK2, MicroScan, Etest, disk diffusion, and broth microdilution. *J Microbiol Methods* 112:87–91. <https://doi.org/10.1016/j.mimet.2015.03.014>. Epub 2015 Mar 17. PMID: 25794901
73. Griffin PM, Price GR, Schooneveldt JM, Schlebusch S, Tilse MH, Urbanski T, Hamilton B, Venter D (2012) Use of matrix-assisted laser desorption ionization-time of flight mass spectrometry to identify vancomycin-resistant enterococci and investigate the epidemiology of an outbreak. *J Clin Microbiol* 50(9):2918–2931. <https://doi.org/10.1128/JCM.01000-12>. Epub 2012 June 27. PMID: 22740710; PMCID: PMC3421795
74. Outhred AC, Jelfs P, Suliman B, Hill-Cawthorne GA, Crawford AB, Marais BJ, Sintchenko V (2015) Added value of whole-genome sequencing for management of highly drug-resistant TB. *J Antimicrob Chemother* 70(4):1198–1202. <https://doi.org/10.1093/jac/dku508>. Epub 2014 Dec 9. PMID: 25492392; PMCID: PMC4356205
75. Dallman TJ, Byrne L, Launders N, Glen K, Grant KA, Jenkins C (2015) The utility and public health implications of PCR and whole genome sequencing for the detection and investigation of an outbreak of Shiga toxin-producing *Escherichia coli* serogroup O26:H11. *Epidemiol Infect* 143(8):1672–1680. <https://doi.org/10.1017/S0950268814002696>. Epub 2014 Oct 15. PMID: 25316375
76. Lowe AM, Beattie DT, Deresiewicz RL (1998) Identification of novel staphylococcal virulence genes by in vivo expression technology. *Mol Microbiol* 27(5):967–976. <https://doi.org/10.1046/j.1365-2958.1998.00741.x>. PMID: 9535086
77. van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, Fussing V, Green J, Feil E, Gerner-Smidt P, Brisse S, Struelens M (2007) European Society of Clinical

- Microbiology and Infectious Diseases (ESCMID) Study Group on Epidemiological Markers (ESGEM). Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clin Microbiol Infect* 13(Suppl 3):1–46. <https://doi.org/10.1111/j.1469-0691.2007.01786.x>. PMID: 17716294
78. Cheung AL, Bayer AS, Zhang G, Gresham H, Xiong YQ (2004) Regulation of virulence determinants in vitro and in vivo in *Staphylococcus aureus*. *FEMS Immunol Med Microbiol* 40(1):1–9. [https://doi.org/10.1016/S0928-8244\(03\)00309-2](https://doi.org/10.1016/S0928-8244(03)00309-2). PMID: 14734180
 79. Saber MM, Shapiro BJ (2020) Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genom* 6(3):e000337. <https://doi.org/10.1099/mgen.0.000337>
 80. Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, Meehan CJ, de Jong B, Rigouts L, Sloutsky A, Kaur D, Sunyaev S, van Soolingen D, Shendure J, Sacchettini J, Murray M (2019) GWAS for quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals resistance genes and regulatory regions. *Nat Commun* 10(1):2128. <https://doi.org/10.1038/s41467-019-10110-6>. PMID: 31086182; PMCID: PMC6513847
 81. Salehizadeh H, Shojaosadati SA (2001) Extracellular biopolymeric flocculants. Recent trends and biotechnological importance. *Biotechnol Adv* 19(5):371–385. [https://doi.org/10.1016/s0734-9750\(01\)00071-4](https://doi.org/10.1016/s0734-9750(01)00071-4). PMID: 14538073
 82. Engvall E, Perlmann P (1971) Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G. *Immunochemistry* 8(9):871–874. [https://doi.org/10.1016/0019-2791\(71\)90454-x](https://doi.org/10.1016/0019-2791(71)90454-x). PMID: 5135623
 83. Kim JK, Adam A, Loo JC, Ong H (1995) A chemiluminescence enzyme immunoassay (CLEIA) for the determination of medroxyprogesterone acetate in human serum. *J Pharm Biomed Anal* 13(7):885–891. [https://doi.org/10.1016/0731-7085\(95\)01503-d](https://doi.org/10.1016/0731-7085(95)01503-d). PMID: 8562612
 84. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature* 422(6928):198–207. <https://doi.org/10.1038/nature01511>. PMID: 12634793
 85. O'Neill LP, Turner BM (2003) Immunoprecipitation of native chromatin: NChIP. *Methods* 31(1):76–82. [https://doi.org/10.1016/s1046-2023\(03\)00090-2](https://doi.org/10.1016/s1046-2023(03)00090-2). PMID: 12893176
 86. Orlando V (2000) Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* 25(3):99–104. [https://doi.org/10.1016/s0968-0004\(99\)01535-2](https://doi.org/10.1016/s0968-0004(99)01535-2). PMID: 10694875
 87. Weinmann AS, Farnham PJ (2002) Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods* 26(1):37–47. [https://doi.org/10.1016/S1046-2023\(02\)00006-3](https://doi.org/10.1016/S1046-2023(02)00006-3). PMID: 12054903
 88. Asp P (2018) How to combine ChIP with qPCR. In: Visa N, Jordán-Pla A (eds) *Chromatin immunoprecipitation. Methods in molecular biology*, vol 1689. Humana, New York. https://doi.org/10.1007/978-1-4939-7380-4_3
 89. Daniel R, Michael G (2003) Genomewide histone acetylation microarrays. *Methods* 31(1):83–89. ISSN 1046-2023. [https://doi.org/10.1016/S1046-2023\(03\)00091-4](https://doi.org/10.1016/S1046-2023(03)00091-4)
 90. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA* 89(5):1827–1831. <https://doi.org/10.1073/pnas.89.5.1827>. PMID: 1542678; PMCID: PMC48546
 91. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 33(18):5868–5877. <https://doi.org/10.1093/nar/gki901>. PMID: 16224102; PMCID: PMC1258174
 92. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* 37(8):853–862. <https://doi.org/10.1038/ng1598>. Epub 2005 July 10. PMID: 16007088

93. Booth MJ, Branco MR, Ficiz G, Oxley D, Krueger F, Reik W, Balasubramanian S (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* 336(6083):934–937. <https://doi.org/10.1126/science.1220671>. Epub 2012 Apr 26. PMID: 22539555
94. Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B, Min JH, Jin P, Ren B, He C (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 149(6):1368–1380. <https://doi.org/10.1016/j.cell.2012.04.027>. Epub 2012 May 17. PMID: 22608086; PMCID: PMC3589129
95. Eid J, Fehr A (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910):133–138. <https://doi.org/10.1126/science.1162986>. Epub 2008 Nov 20. PMID: 19023044
96. Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O’Grady J (2015) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* 33(3):296–300. <https://doi.org/10.1038/nbt.3103>. Epub 2014 Dec 8. PMID: 25485618
97. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 10(10):669–680. <https://doi.org/10.1038/nrg2641>. Epub 2009 Sep 8. PMID: 19736561; PMCID: PMC3191340
98. Rossi E, Cimdins A, Lütjhe P, Brauner A, Sjöling Å, Landini P, Römling U (2018) “It’s a gut feeling”—*Escherichia coli* biofilm formation in the gastrointestinal tract environment. *Crit Rev Microbiol* 44(1):1–30. <https://doi.org/10.1080/1040841X.2017.1303660>. Epub 2017 May 9. Erratum in: *Crit Rev Microbiol* 2018;44(1):i. PMID: 28485690
99. Kim J, Hegde M, Jayaraman A (2010) Co-culture of epithelial cells and bacteria for investigating host-pathogen interactions. *Lab Chip* 10(1):43–50. <https://doi.org/10.1039/b911367c>. Epub 2009 Oct 16. PMID: 20024049
100. Kim HJ, Li H, Collins JJ, Ingber DE (2016) Contributions of microbiome and mechanical deformation to intestinal bacterial overgrowth and inflammation in a human gut-on-a-chip. *Proc Natl Acad Sci USA* 113(1):E7–E15. <https://doi.org/10.1073/pnas.1522193112>. Epub 2015 Dec 14. PMID: 26668389; PMCID: PMC4711860
101. Costello CM, Sorna RM, Goh YL, Cengic I, Jain NK, March JC (2014) 3-D intestinal scaffolds for evaluating the therapeutic potential of probiotics. *Mol Pharm* 11(7):2030–2039. <https://doi.org/10.1021/mp5001422>. Epub 2014 May 13. PMID: 24798584; PMCID: PMC4096232
102. Costello CM, Hongpeng J, Shaffiey S, Yu J, Jain NK, Hackam D, March JC (2014) Synthetic small intestinal scaffolds for improved studies of intestinal differentiation. *Biotechnol Bioeng* 111(6):1222–1232. <https://doi.org/10.1002/bit.25180>. Epub 2014 Jan 22. PMID: 24390638; PMCID: PMC4233677
103. Elzinga J, van der Oost J, de Vos WM, Smidt H (2019) The use of defined microbial communities to model host-microbe interactions in the human gut. *Microbiol Mol Biol Rev* 83(2):e00054-18. <https://doi.org/10.1128/MMBR.00054-18>. PMID: 30867232; PMCID: PMC6684003
104. Schmid CA, Müller A (2013) FoxD3 is a novel, epigenetically regulated tumour suppressor in gastric carcinogenesis. *Gastroenterology* 144(1):22–25. <https://doi.org/10.1053/j.gastro.2012.11.014>. Epub 2012 Nov 16. PMID: 23164571
105. Doherty R, Whiston R, Cormican P, Finlay EK, Coudrey C, Brady C, O’Farrelly C, Meade KG (2016) The CD4(+) T cell methylome contributes to a distinct CD4(+) T cell transcriptional signature in *Mycobacterium bovis*-infected cattle. *Sci Rep* 6:31014. <https://doi.org/10.1038/srep31014>. PMID: 27507428; PMCID: PMC4978967
106. Wang W, Zhou C, Tang H, Yu Y, Zhang Q (2020) Combined analysis of DNA methylome and transcriptome reveal novel candidate genes related to porcine *Escherichia coli* F4ab/ac-Induced Diarrhea. *Front Cell Infect Microbiol* 10:250. <https://doi.org/10.3389/fcimb.2020.00250>. PMID: 32547963; PMCID: PMC7272597

107. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, Jabado OJ, Deikus G, Clark TA, Luong K, Murray IA, Davis BM, Keren-Paz A, Chess A, Roberts RJ, Korlach J, Turner SW, Kumar V, Waldor MK, Schadt EE (2012) Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat Biotechnol* 30(12):1232–1239. <https://doi.org/10.1038/nbt.2432>. Epub 2012 Nov 8. Erratum in: *Nat Biotechnol* 2013; 31(6):566. PMID: 23138224; PMCID: PMC3879109
108. Laszlo AH, Derrington IM, Brinkerhoff H, Langford KW, Nova IC, Samson JM, Bartlett JJ, Pavlenok M, Gundlach JH (2013) Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proc Natl Acad Sci USA* 110(47):18904–18909. <https://doi.org/10.1073/pnas.1310240110>. Epub 2013 Oct 28. PMID: 24167255; PMCID: PMC3839702
109. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W (2017) Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* 14(4):407–410. <https://doi.org/10.1038/nmeth.4184>. Epub 2017 Feb 20. PMID: 28218898
110. McIntyre ABR, Alexander N, Grigorev K, Bezdán D, Sichtig H, Chiu CY, Mason CE (2019) Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. *Nat Commun* 10(1):579. <https://doi.org/10.1038/s41467-019-08289-9>. PMID: 30718479; PMCID: PMC6362088
111. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B (2017) Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods* 14(4):411–413. <https://doi.org/10.1038/nmeth.4189>. Epub 2017 Feb 20. PMID: 28218897; PMCID: PMC5704956
112. Seib KL, Jen FE, Tan A, Scott AL, Kumar R, Power PM, Chen LT, Wu HJ, Wang AH, Hill DM, Luyten YA, Morgan RD, Roberts RJ, Maiden MC, Boitano M, Clark TA, Korlach J, Rao DN, Jennings MP (2015) Specificity of the ModA11, ModA12 and ModD1 epigenetic regulator N(6)-adenine DNA methyltransferases of *Neisseria meningitidis*. *Nucleic Acids Res* 43(8):4150–4162. <https://doi.org/10.1093/nar/gkv219>. Epub 2015 Apr 6. PMID: 25845594; PMCID: PMC4417156
113. Jen FE, Seib KL, Jennings MP (2014) Phasevarions mediate epigenetic regulation of antimicrobial susceptibility in *Neisseria meningitidis*. *Antimicrob Agents Chemother* 58(7):4219–4221. <https://doi.org/10.1128/AAC.00004-14>. Epub 2014 Apr 28. PMID: 24777094; PMCID: PMC4068601
114. Srikhanta YN, Gorrell RJ, Steen JA, Gawthorne JA, Kwok T, Grimmond SM, Robins-Browne RM, Jennings MP (2011) Phasevarion mediated epigenetic gene regulation in *Helicobacter pylori*. *PLoS One* 6(12):e27569. <https://doi.org/10.1371/journal.pone.0027569>. Epub 2011 Dec 5. PMID: 22162751; PMCID: PMC3230613
115. Zaleski P, Wojciechowski M, Piekarczyk A (2005) The role of Dam methylation in phase variation of *Haemophilus influenzae* genes involved in defence against phage infection. *Microbiology (Reading)* 151(Pt 10):3361–3369. <https://doi.org/10.1099/mic.0.28184-0>. PMID: 16207918
116. Anjum A, Brathwaite KJ, Aidley J, Connerton PL, Cummings NJ, Parkhill J, Connerton I, Bayliss CD (2016) Phase variation of a Type IIG restriction-modification enzyme alters site-specific methylation patterns and gene expression in *Campylobacter jejuni* strain NCTC11168. *Nucleic Acids Res* 44(10):4581–4594. <https://doi.org/10.1093/nar/gkw019>. Epub 2016 Jan 18. PMID: 26786317; PMCID: PMC4889913
117. Hu Y, Huang K, An Q, Du G, Hu G, Xue J, Zhu X, Wang CY, Xue Z, Fan G (2016) Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol* 17:88. <https://doi.org/10.1186/s13059-016-0950-z>. PMID: 27150361; PMCID: PMC4858893
118. Tanić M, Beck S (2017) Epigenome-wide association studies for cancer biomarker discovery in circulating cell-free DNA: technical advances and challenges. *Curr Opin Genet Dev* 42:48–55. <https://doi.org/10.1016/j.gde.2017.01.017>. Epub 2017 Feb 16. PMID: 28391083



Genomic Evidence Provides the Understanding of SARS-CoV-2 Composition, Divergence, and Diagnosis

4

Manish Tiwari, Gurparsad Singh Suri, Gurleen Kaur, Baljinder Singh, Sahil Mehta, and Divya Mishra

Abstract

A plethora of studies have shown several types of chromosomal mutations (i.e., deletion, insertion, and substitution) being present for α -CoVs, β -CoVs, γ -CoVs, and δ -CoVs. The current pandemic is caused by the β -CoV and SARS-CoV-2. The attributes of a virus are associated with its genomic composition. Mutations can cause changes in the viral genome that can lead to the crossing of the animal-human barrier or result in a more virulent strain that can increase transmission and pathogenicity for coronaviruses in general and SARS-CoV-2 in particular. Additionally, these mutations may result into new genomic properties. Some mutations have caused changes in the structure of amino acids, and this can be a potential explanation for failure in antiviral therapies. Through genome mapping, we can focus on conserved regions in a viral genome so that even if the virus undergoes a chromosomal mutation into a more virulent strain, the vaccine will still work. One of the prime targets for antiviral therapies against SARS-CoV-2 includes spike protein. Using genomic mapping we can better understand, monitor, and treat a viral infection. With the worldwide spread of Sars-CoV-2, it is ever so important to have a greater understanding of the

M. Tiwari · B. Singh
National Institute of Plant Genome Research, New Delhi, India

G. S. Suri · G. Kaur
California Baptist University, Riverside, CA, USA

S. Mehta
International Center for Genetic Engineering and Biotechnology, New Delhi, India

D. Mishra (✉)
Kansas State University, Manhattan, KS, USA
e-mail: divyam@ksu.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*,
https://doi.org/10.1007/978-981-16-0691-5_4

genomic as well as the proteomic landscapes for the virus, which can guide in the future development of antiviral therapies.

Keywords

Coronavirus · Mutation · Vaccine · Synonymous and non-synonymous substitutions

4.1 Introduction

The global population is facing a crisis that emerged in late December 2019 in Wuhan, China. Patients initially presented cases of pneumonia with an unknown etiology. Various elementary symptoms matching that of flu were reported such as dry cough, sore throat, fever, fatigue, and breathlessness [1]. The causative organism was identified later as an RNA virus belonging to the coronavirus (CoV) family, which already created panic in 2002, as severe acute respiratory syndrome (SARS) and in 2012, as Middle East respiratory syndrome (MERS). The recent member differed from the earlier SARS and MERS and is named as SARS-CoV-2 which caused COVID-19 disease [2]. These are zoonotic viruses which crossed the inter-species barrier and infected humans. This apocalyptic threat posed by the sempiternal virus has ruined the worldwide health system and the economies. The researchers are searching for a solution to mitigate or contain the SARS-CoV-2 virus on a global scale. The approach of next generation sequencing has eased to understand the genomic structure of SARS-CoV-2 and its variants. The SARS-CoV-2 genome was sequenced across the globe at a remarkable speed uncovering the different strains with varying pathogenicity and infectivity arising mainly due to mutation in the genome [3–5]. This chapter deals with understanding the features of coronavirus genomes in general and SARS-CoV-2 in particular along with the pathogenicity, various mutations, and vaccine candidates in development.

4.2 General Etiology and Pathogenicity of Coronaviruses

CoVs are a group of highly diverse viruses which are enclosed in an envelope and possess a positive single-stranded RNA genome of ~29 kb size [6]. They belong to subfamily *Coronavirinae* under the family *Coronaviridae*. These CoVs are representative members of four genera, namely, alpha-CoVs, beta-CoVs, gamma-CoVs, and delta-CoVs. The α - and β -CoVs draw major attention as they are potentiated to cross animal-human barriers and hence are regarded as prime human pathogens [7]. In contrast the γ - and δ -CoVs mainly infect birds but could prove fatal in future if they could also adapt themselves to cross this barrier. Till date seven human coronaviruses (hCoVs) have been reported which infect humans. Two of them belong to α -CoVs genera, hCoV-NL63 and hCoV-229E, and the rest of the five belong to β -CoVs genera, namely, SARS-CoV, MERS-CoV, hCoV-OC43, SARS-

CoV hCoV-HKU1, and recently SARS-CoV-2 [8]. Two α -CoVs, HCoV-NL63 and HCoV-229E, and two β -CoVs, HCoV-OC43 and HCoV-HKU1, possess low pathogenicity and mainly present asymptomatic symptoms such as gastrointestinal infections and common cold, whereas the other three β -CoVs, SARS-CoV, MERS-CoV, and SARS-CoV-2, are highly lethal with dreadful pathogenicity and posed severe life-threatening respiratory tract infections [6]. These three β -CoVs specially attack the lower respiratory tract causing acute respiratory distress syndrome (ARDS), accompanied with most systemic symptoms, multi-organ failure, longest illness durations, and usually high case fatality ratio (CFR) [9]. The first encounter with deadly SARS-CoVs was reported in November 2002 in Foshan city of China which spread globally routed via Hong Kong [10]. The epidemic was finally contained in July 2003 by interruption of the transmission chain in Taiwan [11]. A decade later another dreadful hCoV, MERS-CoV, emerged in Jordan in April 2012 and is still prevalent in Middle East regions till date [12].

4.3 SARS-CoV-2

An epidemic broke out first in Wuhan city of China in December 2019. This disease was caused by a sempiternal virus SARS-CoV-2. The World Health Organization (WHO) warned the international community and announced COVID-19 as a state of emergency to public health on 30 January 2020. Further, as the COVID-19 outbreak spread to more than 100 countries infecting more than 100,000 people, it was declared as a pandemic by WHO on 11 March 2020 [13]. As of 11 October 2020, a total of 37,109,851 cases were reported resulting in 1,070,355 mortalities (~2.9% death rate) (<https://www.who.int/docs/default-source/coronaviruse/situation-reports/20201012-weekly-epi-update-9.pdf>). The American continent had the maximum number of cases 17,794,771 (48% of Total) and deaths 588,867 (55% of Total) followed by Asia and Europe (Table 4.1, Fig. 4.1).

The spread of SARS-CoV-2 occurs through the inhalation of respiratory droplets from an infected person to a healthy individual. The incubation period of SARS-CoV-2 is 2–14 days, and after that the patients start exhibiting disease symptoms. The severity of COVID-19 varied from individuals to individuals, and the pronounced effect has been observed in aged individuals and patients with past comorbidities such as diabetes, hypertension, asthma, obesity, lung, kidney, and liver or heart disorders, immune-compromised patients, etc. [14]

4.4 Mode of Entry and Interaction with Human Immune System

The target cells during the early stages of host-viral interaction involve nasal and bronchial epithelial cells and pneumocytes. This interaction with these cells and SARS-CoV-2 is initiated by the recognition between host angiotensin-converting enzyme 2 (ACE2) receptors and viral spike protein receptor-binding domains

Table 4.1 Newly reported and cumulative COVID-19 confirmed cases and deaths, by WHO region, as of 11 October 2020

WHO region	New cases in the last 7 days (%)	Change in new cases in the last 7 days (%)	Cumulative cases (%)	New deaths in the last 7 days (%)	Change in new deaths in the last 7 days ^a (%)	Cumulative deaths (%)
Americas	804,735 (35%)	6	17,794,771 (48%)	20,509 (52%)	-5	588,867 (55%)
Southeast Asia	575,763 (25%)	-6	7,911,036 (21%)	7750 (20%)	-8	126,917 (12%)
Europe	694,275 (31%)	34	6,918,265 (19%)	6172 (16%)	16	246,709 (23%)
Eastern Mediterranean	138,751 (6%)	10	2,605,478 (7%)	3173 (8%)	13	66,329 (6%)
Africa	29,169 (1%)	11	1,227,719 (3%)	991 (3%)	27	27,255 (3%)
Western Pacific	26,199 (1%)	6	651,841 (2%)	633 (2%)	26	14,265 (1%)
Other	-	-	741 (<1%)	-	-	13 (<1%)
Global	2,268,892 (100%)	10	37,109,851 (100%)	39,228 (100%)	<1	1,070,355 (100%)

Adapted from WHO data

^aPercent change in the number of newly confirmed cases/deaths in the past 7 days, compared to 7 days prior. Regional percentages rounded to the nearest whole number, global totals may not equal 100%

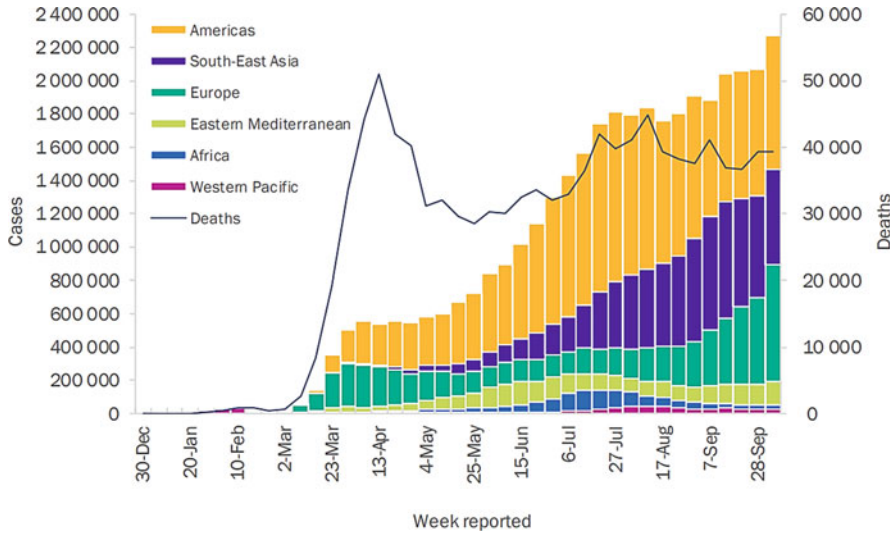


Fig. 4.1 Number of COVID-19 cases reported weekly by WHO region and global deaths, 30 December 2019 through 11 October 2020 (adapted from WHO data)

(RBDs). This is followed by host type 2 transmembrane serine protease (TMPRSS2)-mediated cleavage of ACE2 and activation of SARS-CoV-2 S protein and viral entry into the host [15]. The SARS-CoV-2 infection results in the killing of T lymphocyte cells and lymphopenia. In addition, increased lymphocyte apoptosis and impaired lymphopoiesis are observed associated with the viral inflammatory response. The inflammatory response comprises both the innate and the adaptive immune response [16]. As the viral replication accelerates and infection moves to later stages, the compromised integrity of the epithelial-endothelial barrier is observed. An influx of monocytes and neutrophils is triggered accompanied by an accentuated inflammatory response upon further infection of pulmonary capillary endothelial cells [17].

4.5 Genome Comparison of SARS-CoV-2 with Other CoVs

The genomic length of CoVs ranged between 26 and 32 kb [6, 18]. The MERS-CoV genome is the largest (~30.11 kb), followed by SARS-CoV-2 (~29.9 kb) and SARS-CoV (~27.9 kb) [19]. The alignment of the genomic sequence of SARS-CoV-2 with other α - and β -CoVs revealed that it has low similarity (~65%) with the α -CoVs, whereas similarity increases with the β -CoVs ranging from HCoV-HKU1 (~68%) MERS-CoV (~70) to SARS-CoV (~82%) and SARSr-CoV; RaTG13 (bat-CoV) (~96%) [20, 21]. SARS-CoV-2 shares ~82% similarity with SARS-CoV that is comparatively very low with MERS-CoV (~70%) indicating a closer phylogenetic relatedness between SARS-CoV-2 and SARS-CoV in comparison to MERS-CoV

[22]. Interestingly, the SARS-CoV-2 genome shares maximum similarity to SARS-CoV; RaTG13.

The coronavirus genome may encode a different number (6–11) of open reading frames (ORFs) [23]. The ORF present at the 5' end represents ~67% of the genome and translates into 16 nonstructural proteins. The 3' terminus encodes remaining eight accessory (orf3a, orf3b, orf6, orf7a, orf7b, orf8b, orf9b, and orf14) and four structural proteins (spike protein (S), envelope protein (E), matrix protein (M), and nucleocapsid protein (N)) [24]. The spike protein determines the binding and entry of the virus to the host [23, 25]. Interestingly, the spike proteins of MERS-CoV and SARS-CoV recognize and interact with different host receptors using nonidentical RBDs. On the one hand, MERS-CoV interacts with dipeptidyl peptidase 4 (DPP4, or CD26) receptor [26], whereas SARS-CoV recognizes ACE2 as primary receptors and an alternative CD209L receptor [27, 28]. The SARS-CoV and SARS-CoV-2 possess notable differences at amino acid levels such as the presence of orf8a protein only in SARS-CoV and absent in SARS-CoV-2, a longer version with 121 amino acid of orf8b in SARS-CoV-2, whereas only 84 amino acids in SARS-CoV; contrasting to orf8b the length orf3b protein is longer, 154 amino acids in SARS-CoV compared to only 22 amino acids in SARS-CoV-2 [21] (Table 4.2).

The phylogenetic relationships and amino acid composition of SARS-CoV, MERS-CoV, SARS-like bat-CoVs, and SARS-CoV-2 revealed a close relationship between SARS-CoV-2 and SARS-CoVs or SARS-like bat-CoVs. The genome of the SARS-CoV-2 has several SNPs which results in the change in amino acid and an intervention into the amino acid substitutions present in the proteins of coronaviruses might lead to cues about the structural and functional difference between SARS-CoV-2 and SARS-CoVs.

4.6 Next Generation Sequencing and Identification of Mutation in the SARS-CoV-2 Genome

Next generation sequencing has enormously bolstered the genome sequencing of h-CoV at an unprecedented rate. Till date more than 159,000 h-CoV genomic sequences have been submitted in GISAID database in a short span of ~10 months [67]. These on-the-spot scientific inputs rendered researchers to develop diagnostic kits and revisit the treatment strategies which were better suited and resulted in effective prognosis and containment [68–70]. The genomic sequence has helped researchers globally to identify the various mutating strains. This helped in the prediction and validation of varying pathogenicity of SARS-CoV-2 based on the mutations in the viral genome [3–5, 71]. The genome of SARS-CoV-2 underwent number of mutations over the time due to the interaction with different immune system based on geography and demography. Several investigations revealed that the nucleotide substitution mostly resulted in the incorporation of U nucleotide in the SARS-CoV-2 genome. Kosuge et al. revealed that there is a positive correlation between the increased cytokine inflammatory molecule and percentage of U residue in the SARS-CoV-2 genome [72]. The low abundance of CG dinucleotide in the

Table 4.2 Function of various structural and nonstructural proteins of SARS-CoV-2

Protein	Function	Reference
nsp1	Degradation of cellular mRNA, inhibits interferon signaling and suppresses host gene expression	[29, 30]
nsp2	Suppressed host gene expression in conjunction with nsp1	[31, 32]
nsp3	Protease polypeptides cleaving, augments replication, blocking host innate immune response, promoting cytokine expression	[33, 34]
nsp4	Double-membrane vesicle formation	[35, 36]
nsp5	Chymotrypsin-like protease, polypeptides cleaving, inhibition of interferon signaling	[37, 38]
nsp6	Restricting autophagosome expansion, double-membrane vesicle formation	[39, 40]
nsp7	Cofactor and primase	[41, 42]
nsp8	Cofactor and primase	[41–43]
nsp9	Dimeric RNA-binding protein and responsible for viral infection	[44, 45]
nsp10	Cofactor, scaffold protein for nsp14 and nsp16 and activates the replicating enzyme	[46, 47]
nsp12	Primer-dependent RNA-dependent RNA polymerase	[48, 49]
nsp13	RNA helicase activity and 5' triphosphatase	[50–52]
nsp14	Exoribonuclease	[53, 54]
nsp15	Endoribonuclease	[55, 56]
nsp16	2'-O-methyltransferase activity	[57, 58]
Spike (S)	Receptor binding and viral entry	[59, 60]
Envelope (E)	Viral assembly, release and viral pathogenesis	[61, 62]
Matrix (M)	Viral shape, binds to the nucleocapsid and promotes membrane curvature	[63, 64]
Nucleocapsid (N)	Binds viral genome and augments replication, packaging of the genome in virions	[65, 66]

SARS-CoV-2 genome indicated less energy consumed by the virus to replicate and survive in the host immune system [73]. Both transition and transversion nucleotide substitution events were observed in the SARS-CoV-2 genome [74]. The percentage of transition nucleotide substitution predominates and is higher in the SARS-CoV-2 as compared to the transversion type. These nucleotide substitutions resulted in synonymous and non-synonymous mutations resulting in either similar amino acid or change in amino acid, respectively. The region such as Nsp8-10, Nsp16, and ORF10 is devoid of any non-synonymous mutation [74]. Among the non-synonymous mutations, Spike is one of the most potential candidates which possessed the number of mutations and resulted in the change in amino acid such as D614G, S943P, L5F, L8V, G476S, V483A, V367F, Q239K, A831V, D839Y, and P1263L [71, 75]. D614G is one of the common mutations in the spike region [71, 76, 77]. The frequency of D614 form is more within the population in early pandemic time after that the 614G form is more prevalent now over the time especially in Europe. Korber et al. proposed two hypotheses about the D614G mutation and its effect on the viral infectivity. In the first hypothesis, the mutation disrupts the

interaction between the structural unit of spike protein S1 and S2; hence, the S1 unit was unable to bound with the membrane-bound S2 unit [75]. Another hypothesis is that D614G may affect the binding of receptor-binding domain and ACE2 receptor. The strong interaction of 614G form of spike protein with TMPRSS2 protease could be the reason for loss of interaction between structural units S1 and S2 [78]. The mutation such as S943P and V367F in the spike region is localized to European population [5, 78, 79]. Bioinformatic investigations showed that mutation V367F is responsible for maintaining the stability of receptor-binding domain of spike protein. L5F is localized to European and American population, while L8V is localized to Asian population [71, 80, 81]. Although it is difficult to state but warrants an urgent insights into the implication of these mutations in the viral replication. Additionally, few mutations such as G476S and V483A are found in the American continents [82, 83]. Structural unit S1 possessed H49Y and Q239K, while S2 possessed A831V and D839Y in the spike region of SARS-CoV-2 of local population [75].

Apart from the mutation in the spike region, the other regions such as ORF3a, ORF8, and non-structural proteins Nsp2, Nsp12, and Nsp13 also possessed the mutations [5, 74, 84]. An investigation carried by Yang et al. revealed that 11 mutations are frequent in the SARS-CoV-2 regardless of the demography and geography [84]. These mutations alter the pathogenicity and infectivity of viral strains. Characterizing these mutations in future may hold the key to better understand the viral interaction with the host and identification of weak strains which can be used for inoculation to provide immunity against coronavirus [74].

4.7 Vaccine Development

One of the key goals of SARS-CoV-2 research globally is to contain and mitigate this virus by developing an effective antiviral therapy in form of drugs or vaccines. Till 15 October 2020, 42 candidate vaccines were in clinical evaluation, out of which 10 were in Phase 3 trial (<https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines>) (Table 4.3). Most of the vaccine candidates are composed of inactivated virus, RNA molecules, protein subunits, DNA molecules, and non-replicating viral vector. The spike protein has been considered as the prime target of the vaccine formulations against SARS-CoV-2, and the recent mapping of the spike protein may provide insights for an expedited development of a specific vaccine [85].

The world is preparing for the influenza virus before emergence of SARS-CoV-2. Antibody responses against the viruses such as influenza and SARS-CoV-2 can provide immunity towards homologous viruses, and the efficacy declined considerably against antigenic variants and failed miserably in countering genetically shifted viruses. Traditional vaccines were formulated using pathogen-derived purified proteins or attenuated virus, and they stimulate antibodies production. Although it is an effective immunization technique, the process can take years. The time frame for production and immunization of killed or attenuated virus vaccines requires a minimum of 6 months since strain identification to distribution of vaccine [86]. This

Table 4.3 Vaccine candidates based on different platforms under clinical trial

COVID-19 vaccine developer/manufacturer	Vaccine platform	Type of candidate vaccine	Route of Administration
Sinovac	Inactivated	Inactivated	IM
Wuhan Institute of Biological Products/Sinopharm	Inactivated	Inactivated	IM
Beijing Institute of Biological Products/Sinopharm	Inactivated	Inactivated	IM
University of Oxford/AstraZeneca	Non-replicating viral vector	ChAdOx1-S	IM
CanSino Biological Inc./Beijing Institute of Biotechnology	Non-replicating viral vector	Adenovirus type 5 vector	IM
Gamaleya Research Institute	Non-replicating viral vector	Adeno-based (rAd26-S+rAd5-S)	IM
Janssen Pharmaceutical Companies	Non-replicating viral vector	Ad26COVS1	IM
Novavax	Protein subunit	Full-length recombinant SARS CoV-2 glycoprotein nanoparticle vaccine adjuvanted with matrix M	IM
Moderna/NIAD	RNA	LNP-encapsulated mRNA	IM
BioNTech/Fosun Pharma/Pfizer	RNA	3 LNP-mRNAs	IM
Anhui Zhifei Longcom Biopharmaceutical/Institute of Microbiology, Chinese Academy of Sciences	Protein subunit	Adjuvanted recombinant protein (RBD-dimer)	IM
Curevac	RNA	mRNA	IM
Institute of Medical Biology, Chinese Academy of Medical Sciences	Inactivated	Inactivated	IM
Research Institute for Biological Safety Problems, Rep of Kazakhstan	Inactivated	Inactivated	IM
Inovio Pharmaceuticals/International Vaccine Institute	DNA	DNA plasmid vaccine with electroporation	ID
Osaka University/AnGes/Takara Bio	DNA	DNA plasmid vaccine + Adjuvant	IM
Cadila Healthcare Limited	DNA	DNA plasmid vaccine	ID
Genexine Consortium	DNA	DNA vaccine (GX-19)	IM
Bharat Biotech	Inactivated	Whole-Virion Inactivated	IM
Kentucky Bioprocessing, Inc	Protein subunit	RBD-based	IM

(continued)

Table 4.3 (continued)

COVID-19 vaccine developer/manufacturer	Vaccine platform	Type of candidate vaccine	Route of Administration
Sanofi Pasteur/GSK	Protein subunit	S protein (baculovirus production)	IM
Arcturus/Duke-NUS	RNA	mRNA	IM
SpyBiotech/Serum Institute of India	VLP	RBD-HBsAg VLPs	IM
Beijing Minhai Biotechnology Co., Ltd.	Inactivated	Inactivated	IM
ReiThera/LEUKOCARE/Univercells	Non-replicating viral vector	Replication defective Simian adenovirus (GRAd) encoding S	IM
CanSino Biological Inc./Institute of Biotechnology, Academy of Military Medical Sciences, PLA of China	Non-replicating viral vector	Ad5-nCoV	IM/mucosal
Vaxart	Non-replicating viral vector	Ad5 adjuvanted oral vaccine platform	Oral
Ludwig Maximilian University of Munich	Non-replicating viral vector	MVA-SARS-2-S	IM
Clover Biopharmaceuticals Inc./GSK/Dynavax	Protein subunit	Native-like trimeric subunit spike protein vaccine	IM
Vaxine Pty Ltd/Medytox	Protein subunit	Recombinant spike protein with Advax™ adjuvant	IM
University of Queensland/CSL/Seqirus	Protein subunit	Molecular clamp stabilized spike protein with MF59 adjuvant	IM
Medigen Vaccine Biologics Corporation/NIAID/Dynavax	Protein subunit	S-2P protein + CpG 1018	IM
Instituto Finlay de Vacunas, Cuba	Protein subunit	RBD + adjuvant	IM
FBRI SRC VB VECTOR, Rospotrebnadzor, Koltsovo	Protein subunit	Peptide	IM
West China Hospital, Sichuan University	Protein subunit	RBD (baculovirus production expressed in Sf9 cells)	IM
University Hospital Tuebingen	Protein subunit	SARS-CoV-2 HLA-DR peptides	SC
COVAXX	Protein subunit	S1-RBD protein	IM
Institute Pasteur/Themis/Univ. of Pittsburgh CVR/Merck Sharp & Dohme	Replicating viral vector	Measles vector-based	IM

(continued)

Table 4.3 (continued)

COVID-19 vaccine developer/manufacturer	Vaccine platform	Type of candidate vaccine	Route of Administration
Beijing Wantai Biological Pharmacy/Xiamen University	Replicating viral vector	Intranasal flu-based RBD	IM
Imperial College London	RNA	LNP-nCoVsaRNA	IM
People's Liberation Army (PLA) Academy of Military Sciences/Walvax Biotechnology	RNA	mRNA	IM
Medicago Inc.	VLP	Plant-derived VLP adjuvanted with GSK or Dynavax adjs.	IM

time period is an issue of concern during a global pandemic such as SARS-CoV-2 [87]. The pandemics demand development of novel vaccine strategies that can reduce production times and can simultaneously answer for genetic drift inducing broad-spectrum immunity against mutating viral strains. The solution to these concerns could be largely explained by the use of nucleic acid vaccines (DNA and RNA vaccines) as they can be designed in short time span against any viral sequence [86]. Importantly, the nucleic acid vaccines can be designed against more conserved antigen sequences and hence can yield a universal vaccine solution to genetic drifts and unknown future pandemics. In view of this, a new vaccine platform has been developed in form of RNA vaccines. The RNA vaccines can potentially elicit immune responses against several cancers and infectious agents such as bacteria and viruses [87, 88].

A combinatorial use of reverse genetics approach and next-generation sequencing may positively impact the development time of conventional vaccines during epidemics [86].

4.8 Conclusions and Future Directives

Considering the current tally of more than 44 million being already infected since the onset of catastrophic readily transmissible COVID-19, various researchers around the globe have generated a lot of knowledge on the pathogenesis, fatality rate, infectivity rate, genomic organization, and variations. Apart from the efforts done by the scientific community in open reporting and sharing data related to COVID-19, various concerned governing authorities have also played their part well. Their efforts included raising awareness about this contagious virus, providing time-to-time guidelines, approving containment zones, necessary lockdowns, quarantining infected people, and supporting the development of rapid and sensitive diagnostic kits. Nevertheless, to support the governmental bodies to maximum, the health-related workers made an impact by devoting both time and efforts to treating the

SARS-CoV-2 ill patients with available FDA approved repository of antiviral antibodies, vaccines, and drugs that were developed for previously emerged hCOVs. The list included chloroquine, hydroxychloroquine, lopinavir, ritonavir, tocilizumab, and angiotensin receptor blocker, however, a few of them were not effective in curing the disease in a larger cohort. All this has been done as it was the need of the hour to break the chain of community spread by the SARS-CoV-2 pandemic.

The tally for the number of infections and mortalities increased exponentially over time that necessitated the academic laboratories and industry researchers to move at a rapid pace. Together, these researchers provided hope for success by developing many viable vaccine candidates that are currently being evaluated in many countries. This can be supported by the fact that within 40 days of initial efforts, the first vaccine candidate was ready to enter the clinical development pipeline. The R&D processes for vaccine development are being done in this critical period at a global scale so as to minimize the possibility of heavy losses associated with the probable second wave of COVID-19 that will hit by winters. Given the imperative for vaccine development, there are indications from developers, distributors, policymakers, and regulating bodies that vaccines may be available by early 2021 because many viable candidates have already shown higher efficacy in Phase 3. This has been achievable because of the combination of urgent necessity, sequencing-support, pre-available literature on coronaviruses, innovative regulatory processes, scaling manufacturing capacity coupled with an unprecedented shift from the traditional as well as accelerated vaccine development pathways in a short span of time.

However, first and foremost, there are still beads of more questions than answers around etiology, disease mechanisms, immunoregulation (cellular and humoral immune responses) during SARS-CoV-2 that will be addressed by basic research and umpteen experimental trials in the future to bring the human health back on track.

In near future, a globally accessible, FDA-approved vaccine could be developed after laying a solid foundation of research as per scientific laws. After development, the vaccine will be manufactured in sufficient quantities followed by equitable supply to every region of the world including low-resource, highly affected areas. To facilitate the above-described statement, a strong network of international coordination between all the pillars related to vaccine regulation, manufacturers, funding, distribution need to be built to ensure the promise of a global vaccination program for developing a “COVID-19 free world.”

References

1. Shah VK, Fimal P, Alam A, Ganguly D, Chattopadhyay S (2020) Overview of immune response during SARS-CoV-2 infection: lessons from the past. *Front Immunol* 11:1949. Available: <https://www.frontiersin.org/article/10.3389/fimmu.2020.01949>

2. Wang L, Wang Y, Ye D, Liu Q (2020) International Journal of Antimicrobial Agents Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence. *Int J Antimicrob Agents* 55:105948. <https://doi.org/10.1016/j.ijantimicag.2020.105948>
3. Mishra D, Suri GS, Kaur G, Tiwari M (2020) A comparative insight into genomic landscape of SARS-CoV-2 and identification of mutations associated with origin of infection and diversity. *J Med Virol*. <https://doi.org/10.1002/jmv.26744>
4. Saha DP, Majumder R, Chakraborty S, Srivastava AK, Mandal M, Sarkar S (2020) Mutations in spike protein of SARS-CoV-2 modulate receptor binding, membrane fusion and immunogenicity: an insight into viral tropism and pathogenesis of COVID-19. *ChemRxiv*. <https://doi.org/10.26434/chemrxiv.12320567.v1>
5. Mercatelli D, Giorgi FM (2020) Geographic and genomic distribution of SARS-CoV-2 mutations. *Front Microbiol*. <https://doi.org/10.3389/fmicb.2020.01800>
6. Weiss SR (2020) Forty years with coronaviruses. *J Exp Med* 217. <https://doi.org/10.1084/jem.20200537>
7. Coleman CM, Frieman MB (2014) Coronaviruses: important emerging human pathogens. *J Virol* 88:5209–5212. <https://doi.org/10.1128/JVI.03488-13>
8. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA et al (2020) The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* 5:536–544. <https://doi.org/10.1038/s41564-020-0695-z>
9. Vos LM, Bruyndonckx R, Zuithoff NPA, Little P, Oosterheert JJ, Broekhuizen BDL et al (2020) Lower respiratory tract infection in the community: associations between viral aetiology and illness course. *Clin Microbiol Infect*. <https://doi.org/10.1016/j.cmi.2020.03.023>
10. Hui DSC, Zumla A (2019) Severe acute respiratory syndrome: historical, epidemiologic, and clinical features. *Infect Dis Clin North Am* 33:869–889. <https://doi.org/10.1016/j.idc.2019.07.001>
11. Zhu Z, Lian X, Su X, Wu W, Marraro GA, Zeng Y (2020) From SARS and MERS to COVID-19: a brief summary and comparison of severe acute respiratory infections caused by three highly pathogenic human coronaviruses. *Respir Res* 21:224. <https://doi.org/10.1186/s12931-020-01479-w>
12. Hijawi B, Abdallat M, Sayaydeh A, Alqasrawi S, Haddadin A, Jaarour N et al (2013) Novel coronavirus infections in Jordan, April 2012: epidemiological findings from a retrospective investigation. *East Mediterr Heal J (La Rev sante la Mediterr Orient: al-Majallah al-sihhiyah li-sharq al-mutawassit)* 19(suppl 1):S12–S18
13. WHO (2020) https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf?sfvrsn=1ba62e57_10
14. Jordan RE, Adab P, Cheng KK (2020) Covid-19: risk factors for severe disease and death. *BMJ* 368:m1198. <https://doi.org/10.1136/bmj.m1198>
15. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S et al (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* 181:271–280.e8. <https://doi.org/10.1016/j.cell.2020.02.052>
16. Fosbøl EL, Butt JH, Østergaard L, Andersson C, Selmer C, Kragholm K et al (2020) Association of angiotensin-converting enzyme inhibitor or angiotensin receptor blocker use with COVID-19 diagnosis and mortality. *JAMA* 324:168–177. <https://doi.org/10.1001/jama.2020.11301>
17. Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C et al (2020) Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *Lancet Respir Med* 420–422. [https://doi.org/10.1016/S2213-2600\(20\)30076-X](https://doi.org/10.1016/S2213-2600(20)30076-X)
18. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J et al (2016) Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol* 24:490–502. <https://doi.org/10.1016/j.tim.2016.03.003>
19. Al-Qahtani AA (2020) Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): emergence, history, basic and clinical aspects. *Saudi J Biol Sci* 27:2531–2538. <https://doi.org/10.1016/j.sjbs.2020.04.033>

20. Kaur N, Singh R, Dar Z, Bijarnia RK, Dhingra N, Kaur T (2020) Genetic comparison among various coronavirus strains for the identification of potential vaccine targets of SARS-CoV2. *Infect Genet Evol.* <https://doi.org/10.1016/j.meegid.2020.104490>
21. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P et al (2020) Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe* 27:325–328. <https://doi.org/10.1016/j.chom.2020.02.001>
22. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H et al (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395:565–574. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8)
23. Song Z, Xu Y, Bao L, Zhang L, Yu P, Qu Y et al (2019) From SARS to MERS, thrusting coronaviruses into the spotlight. *Viruses.* <https://doi.org/10.3390/v11010059>
24. Cui J, Li F, Shi Z-L (2019) Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* 17:181–192
25. Li F (2016) Structure, function, and evolution of coronavirus spike proteins. *Annu Rev Virol* 3:237–261. <https://doi.org/10.1146/annurev-virology-110615-042301>
26. Li Y, Zhang Z, Yang L, Lian X, Xie Y, Li S et al (2020) The MERS-CoV receptor DPP4 as a candidate binding target of the SARS-CoV-2 Spike. *iScience* 23:101160. <https://doi.org/10.1016/j.isci.2020.101160>
27. Ge X-Y, Li J-L, Yang X-L, Chmura AA, Zhu G, Epstein JH et al (2013) Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* 503:535–538
28. Jeffers SA, Tusell SM, Gillim-Ross L, Hemmila EM, Achenbach JE, Babcock GJ et al (2004) CD209L (L-SIGN) is a receptor for severe acute respiratory syndrome coronavirus. *Proc Natl Acad Sci USA* 101:15748–15753
29. Huang C, Lokugamage KG, Rozovics JM, Narayanan K, Semler BL, Makino S (2011) SARS coronavirus nsp1 protein induces template-dependent endonucleolytic cleavage of mRNAs: viral mRNAs are resistant to nsp1-induced RNA cleavage. *PLoS Pathog* 7:e1002433. Available: <https://doi.org/10.1371/journal.ppat.1002433>
30. Tanaka T, Kamitani W, DeDiego ML, Enjuanes L, Matsuura Y (2012) Severe acute respiratory syndrome coronavirus nsp1 facilitates efficient propagation in cells through a specific translational shutoff of host mRNA. *J Virol* 86:11128–11137. <https://doi.org/10.1128/JVI.01700-12>
31. Graham RL, Sims AC, Brockway SM, Baric RS, Denison MR (2005) The nsp2 replicase proteins of murine hepatitis virus and severe acute respiratory syndrome coronavirus are dispensable for viral replication. *J Virol* 79:13399–13411. <https://doi.org/10.1128/JVI.79.21.13399-13411.2005>
32. Gadlage MJ, Graham RL, Denison MR (2008) Murine coronaviruses encoding nsp2 at different genomic loci have altered replication, protein expression, and localization. *J Virol* 82:11964–11969. <https://doi.org/10.1128/JVI.01126-07>
33. Lei J, Kusov Y, Hilgenfeld R (2018) Nsp3 of coronaviruses: structures and functions of a large multi-domain protein. *Antiviral Res* 149:58–74. <https://doi.org/10.1016/j.antiviral.2017.11.001>
34. Serrano P, Johnson MA, Chatterjee A, Neuman BW, Joseph JS, Buchmeier MJ et al (2009) Nuclear magnetic resonance structure of the nucleic acid-binding domain of severe acute respiratory syndrome coronavirus nonstructural protein 3. *J Virol* 83:12998–13008. <https://doi.org/10.1128/JVI.01253-09>
35. Beachboard DC, Anderson-Daniels JM, Denison MR (2015) Mutations across murine hepatitis virus nsp4 alter virus fitness and membrane modifications. Perlman S (ed). *J Virol* 89:2080–2089. <https://doi.org/10.1128/JVI.02776-14>
36. Gadlage MJ, Sparks JS, Beachboard DC, Cox RG, Doyle JD, Stobart CC et al (2010) Murine hepatitis virus nonstructural protein 4 regulates virus-induced membrane modifications and replication complex function. *J Virol* 84:280–290. <https://doi.org/10.1128/JVI.01772-09>
37. Zhu X, Fang L, Wang D, Yang Y, Chen J, Ye X et al (2017) Porcine deltacoronavirus nsp5 inhibits interferon- β production through the cleavage of NEMO. *Virology* 502:33–38. <https://doi.org/10.1016/j.virol.2016.12.005>

38. Stobart CC, Sexton NR, Munjal H, Lu X, Molland KL, Tomar S et al (2013) Chimeric exchange of coronavirus nsp5 proteases (3CLpro) identifies common and divergent regulatory determinants of protease activity. *J Virol* 87:12611–12618. <https://doi.org/10.1128/JVI.02050-13>
39. Angelini MM, Akhlaghpour M, Neuman BW, Buchmeier MJ (2013) Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. Moscona A (ed). *MBio* 4:e00524–e00513. <https://doi.org/10.1128/mBio.00524-13>
40. Cottam EM, Whelband MC, Wileman T (2014) Coronavirus NSP6 restricts autophagosome expansion. *Autophagy* 10:1426–1441. <https://doi.org/10.4161/autophagy.29309>
41. Kirchdoerfer RN, Ward AB (2019) Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat Commun* 10:2342. <https://doi.org/10.1038/s41467-019-10280-3>
42. Zhai Y, Sun F, Li X, Pang H, Xu X, Bartlam M et al (2005) Insights into SARS-CoV transcription and replication from the structure of the nsp7–nsp8 hexadecamer. *Nat Struct Mol Biol* 12:980–986. <https://doi.org/10.1038/nsmb999>
43. te Velthuis AJW, van den Worm SHE, Snijder EJ (2012) The SARS-coronavirus nsp7+nsp8 complex is a unique multimeric RNA polymerase capable of both de novo initiation and primer extension. *Nucleic Acids Res* 40:1737–1747. <https://doi.org/10.1093/nar/gkr893>
44. Egloff M-P, Ferron F, Campanacci V, Longhi S, Rancurel C, Dutartre H et al (2004) The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world. *Proc Natl Acad Sci USA* 101:3792–3796. <https://doi.org/10.1073/pnas.0307877101>
45. Zeng Z, Deng F, Shi K, Ye G, Wang G, Fang L et al (2018) Dimerization of coronavirus nsp9 with diverse modes enhances its nucleic acid binding affinity. Gallagher T (ed). *J Virol* 92:e00692–e00618. <https://doi.org/10.1128/JVI.00692-18>
46. Bouvet M, Lugari A, Posthuma CC, Zevenhoven JC, Bernard S, Betzi S et al (2014) Coronavirus Nsp10, a critical co-factor for activation of multiple replicative enzymes. *J Biol Chem* 289:25783–25796. <https://doi.org/10.1074/jbc.M114.577353>
47. Ma Y, Wu L, Shaw N, Gao Y, Wang J, Sun Y et al (2015) Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. *Proc Natl Acad Sci USA* 112:9436–9441. <https://doi.org/10.1073/pnas.1508686112>
48. Ahn D-G, Choi J-K, Taylor DR, Oh J-W (2012) Biochemical characterization of a recombinant SARS coronavirus nsp12 RNA-dependent RNA polymerase capable of copying viral RNA templates. *Arch Virol* 157:2095–2104. <https://doi.org/10.1007/s00705-012-1404-x>
49. te Velthuis AJW, Arnold JJ, Cameron CE, van den Worm SHE, Snijder EJ (2010) The RNA polymerase activity of SARS-coronavirus nsp12 is primer dependent. *Nucleic Acids Res* 38:203–214. <https://doi.org/10.1093/nar/gkp904>
50. Adedeji AO, Lazarus H (2016) Biochemical characterization of middle east respiratory syndrome coronavirus helicase. Frieman MB (ed). *mSphere* 1:e00235–e00216. <https://doi.org/10.1128/mSphere.00235-16>
51. Hao W, Wojdyla JA, Zhao R, Han R, Das R, Zlatev I et al (2017) Crystal structure of Middle East respiratory syndrome coronavirus helicase. *PLoS Pathog* 13:e1006474. Available: <https://doi.org/10.1371/journal.ppat.1006474>
52. Jia Z, Yan L, Ren Z, Wu L, Wang J, Guo J et al (2019) Delicate structural coordination of the severe acute respiratory syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res* 47:6538–6550. <https://doi.org/10.1093/nar/gkz409>
53. Bouvet M, Imbert I, Subissi L, Gluais L, Canard B, Decroly E (2012) RNA 3'-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. *Proc Natl Acad Sci USA* 109:9372–9377. <https://doi.org/10.1073/pnas.1201130109>
54. Chen Y, Cai H, Pan J, Xiang N, Tien P, Ahola T et al (2009) Functional screen reveals SARS coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase. *Proc Natl Acad Sci USA* 106:3484–3489. <https://doi.org/10.1073/pnas.0808790106>

55. Deng X, Hackbart M, Mettelman RC, O'Brien A, Mielech AM, Yi G et al (2017) Coronavirus nonstructural protein 15 mediates evasion of dsRNA sensors and limits apoptosis in macrophages. *Proc Natl Acad Sci USA* 114:E4251–E4260. <https://doi.org/10.1073/pnas.1618310114>
56. Zhang L, Li L, Yan L, Ming Z, Jia Z, Lou Z et al (2018) Structural and biochemical characterization of endoribonuclease Nsp15 Encoded by Middle East respiratory syndrome coronavirus. Gallagher T (ed). *J Virol* 92:e00893–e00818. <https://doi.org/10.1128/JVI.00893-18>
57. Snijder EJ, Decroly E, Ziebuhr J (2016) The nonstructural proteins directing coronavirus RNA synthesis and processing. *Adv Virus Res* 96:59–126. <https://doi.org/10.1016/bs.aivir.2016.08.008>
58. Shi P, Su Y, Li R, Liang Z, Dong S, Huang J (2019) PEDV nsp16 negatively regulates innate immunity to promote viral proliferation. *Virus Res* 265:57–66. <https://doi.org/10.1016/j.virusres.2019.03.005>
59. Delmas B, Laude H (1990) Assembly of coronavirus spike protein into trimers and its role in epitope expression. *J Virol* 64:5367–5375
60. Beniac DR, Andonov A, Grudski E, Booth TF (2006) Architecture of the SARS coronavirus prefusion spike. *Nat Struct Mol Biol* 13:751–752
61. Nieto-Torres JL, DeDiego ML, Verdiá-Báguena C, Jimenez-Guardeño JM, Regla-Nava JA, Fernandez-Delgado R et al (2014) Severe acute respiratory syndrome coronavirus envelope protein ion channel activity promotes virus fitness and pathogenesis. *PLoS Pathog* 10:e1004077
62. DeDiego ML, Álvarez E, Almazán F, Rejas MT, Lamirande E, Roberts A et al (2007) A severe acute respiratory syndrome coronavirus that lacks the E gene is attenuated in vitro and in vivo. *J Virol* 81:1701–1713
63. Neuman BW, Kiss G, Kunding AH, Bhella D, Baksh MF, Connelly S et al (2011) A structural analysis of M protein in coronavirus assembly and morphology. *J Struct Biol* 174:11–22
64. Nal B, Chan C, Kien F, Siu L, Tse J, Chu K et al (2005) Differential maturation and subcellular localization of severe acute respiratory syndrome coronavirus surface proteins S, M and E. *J Gen Virol* 86:1423–1434
65. Cui L, Wang H, Ji Y, Yang J, Xu S, Huang X et al (2015) The nucleocapsid protein of coronaviruses acts as a viral suppressor of RNA silencing in mammalian cells. *J Virol* 89:9029–9043
66. Fehr AR, Perlman S (2015) Coronaviruses: an overview of their replication and pathogenesis. In: *Coronaviruses*. Springer, New York, pp 1–23
67. Elbe S, Buckland-Merrett G (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Challenges* 1:33–46. <https://doi.org/10.1002/gch2.1018>
68. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q (2020) Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* 367:1444–1448. <https://doi.org/10.1126/science.abb2762>
69. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L et al (2020) Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α -ketoamide inhibitors. *Science* 368:409–412. <https://doi.org/10.1126/science.abb3405>
70. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G et al (2020) A new coronavirus associated with human respiratory disease in China. *Nature* 579:265–269. <https://doi.org/10.1038/s41586-020-2008-3>
71. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W et al (2020) Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 182:812–827.e19. <https://doi.org/10.1016/j.cell.2020.06.043>
72. Kosuge M, Furusawa-Nishii E, Ito K, Saito Y (2020) Point mutation bias in SARS-CoV-2 variants results in increased ability to stimulate inflammatory responses. *Sci Rep* 10(1):17766
73. Wang Y, Mao J-M, Wang G-D, Luo Z-P, Yang L, Yao Q et al (2020) Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames. *Sci Rep* 10:12331. <https://doi.org/10.1038/s41598-020-69342-y>

74. Tiwari M, Mishra D (2020) Investigating the genomic landscape of novel coronavirus (2019-nCoV) to identify non-synonymous mutations for use in diagnosis and drug design. *J Clin Virol*. <https://doi.org/10.1016/j.jcv.2020.104441>
75. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W et al (2020) Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv*. <https://doi.org/10.1101/2020.04.29.069054>
76. Kleine-Weber H, Elzayat MT, Wang L, Graham BS, Müller MA, Drosten C et al (2019) Mutations in the spike protein of Middle East respiratory syndrome coronavirus transmitted in Korea increase resistance to antibody-mediated neutralization. *J Virol* 93. <https://doi.org/10.1128/JVI.01381-18>
77. Ogawa J, Zhu W, Tonnu N, Singer O, Hunter T, Ryan (Firth) AL et al (2020) The D614G mutation in the SARS-CoV2 spike protein increases infectivity in an ACE2 receptor dependent manner. *bioRxiv*. <https://doi.org/10.1101/2020.07.21.214932>
78. Raghav S, Ghosh A, Turuk J, Kumar S, Jha A, Madhulika S et al (2020) SARS-CoV2 genome analysis of Indian isolates and molecular modelling of D614G mutated spike protein with TMPRSS2 depicted its enhanced interaction and virus infectivity. *bioRxiv*. <https://doi.org/10.1101/2020.07.23.217430>
79. Banerjee AK, Begum F, Ray U (2020) Mutation hot spots in spike protein of COVID-19. *Preprints*. <https://doi.org/10.20944/preprints202004.0281.v1>
80. Mishra S (2020) Designing of cytotoxic and helper T cell epitope map provides insights into the highly contagious nature of the pandemic novel coronavirus SARS-CoV-2. *R Soc Open Sci* 7:201141. <https://doi.org/10.1098/rsos.201141>
81. Wang R, Hozumi Y, Yin C, Wei G-W (2020) Decoding SARS-CoV-2 transmission and evolution and ramifications for COVID-19 diagnosis, vaccine, and medicine. *J Chem Inf Model*. <https://doi.org/10.1021/acs.jcim.0c00501>
82. Yellapu NK, Patel S, Zhang B, Meier R, Neums L, Pei D et al (2020) Evolutionary analysis of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) reveals genomic divergence with implications for universal vaccine efficacy. *Vaccines*. <https://doi.org/10.3390/vaccines8040591>
83. Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP (2020) The trinity of COVID-19: immunity, inflammation and intervention. *Nat Rev Immunol* 20:363–374. <https://doi.org/10.1038/s41577-020-0311-8>
84. Yang H-C, Chen C, Wang J-H, Liao H-C, Yang C-T, Chen C-W et al (2020) Genomic, geographic and temporal distributions of SARS-CoV-2 mutations. *bioRxiv*. <https://doi.org/10.1101/2020.04.22.055863>
85. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O et al (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 367:1260–1263. <https://doi.org/10.1126/science.abb2507>
86. Lurie N, Saville M, Hatchett R, Halton J (2020) Developing Covid-19 vaccines at pandemic speed. *N Engl J Med* 382:1969–1973. <https://doi.org/10.1056/NEJMp2005630>
87. Erasmus JH, Fuller DH (2020) Preparing for pandemics: RNA vaccines at the forefront. *Mol Ther* 28:1559–1560. <https://doi.org/10.1016/j.ymthe.2020.06.017>
88. Fuller DH, Berglund P (2020) Amplifying RNA vaccine development. *N Engl J Med* 382:2469–2471. <https://doi.org/10.1056/NEJMcibr2009737>



Importance of Next-Generation Sequencing in Viral Diagnostics

5

Ashish Kumar Vyas and Sudheer Gupta

Abstract

The use of real-time PCR has become gold standard for many of the viral diagnosis at present and has replaced the need for culture-based and immunofluorescence-based techniques because of being relatively less time-consuming and inexpensive. Availability of whole genome sequences of viruses accelerated the development of quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) assays for diagnosis of viruses. Next-generation sequencing (NGS) technologies have emerged as powerful technique to perform massive parallel sequencing which enables us to sequence large number of samples with almost no limitation of pathogenic agents and targets on the genome. Application of NGS in viral diagnosis includes viral discovery using metagenomic approach and targeted sequencing approach. However, this current technology has many limitations as well. In this chapter, we summarized the role of NGS in viral diagnostics, applications, and limitations.

Keywords

Next-generation sequencing · Viral diagnostics · Polymerase chain reaction · Viruses

A. K. Vyas (✉)

Department of Microbiology, All India Institute of Medical Sciences (AIIMS), Bhopal, India
e-mail: ashishkvyas.microbiology@aiimsbhopal.edu

S. Gupta

Regional Virology Laboratory, All India Institute of Medical Sciences (AIIMS), Bhopal, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_5

5.1 Introduction

Among the modalities for viral diagnosis, ELISA and real-time PCR have been the most accepted technologies. Serological methods like ELISA (IgM and antigen) suffer from limitation of detectability in the initial days of illness (IgM) and cross-reactivity (antigen) [1]. The use of real-time PCR has become gold standard for many of the viral diagnosis at present and has replaced the need for culture-based and immunofluorescence-based techniques because of being relatively less time-consuming and inexpensive. Further, molecular methods like real-time PCR can also be developed in less than no time for novel and emerging viruses provided the genomic sequences for the virus is known [2]. This requirement of prior knowledge of the viral sequences renders this technique limiting for novel pathogen and less suitable for identification of novel variants. Furthermore, since there is a limit of targets for pathogen, not many pathogens can be detected in a single go in a multiplexed test kit in real-time PCR. However, any syndrome, where diagnosis algorithm is based on symptoms, requires detection of large panel of pathogens, for example, AES (JE, DENV, CHIKV, WNV, herpesvirus, and *Enterovirus*) [3], fever with rash (measles, rubella, DENV, and CHIKV), SARI/ILI [4] (influenza A/B, parainfluenza, adenovirus, RSV, and *Metapneumovirus*), and viral hepatitis (HBV, HCV, HAV and HEV). A real-time PCR-based panel with optimal sensitivity would require large number of multiplexing and tubes for including the panel of viruses, which become extremely tedious when the number of samples is high. Moreover, the testing panel further required regular revisions because of the frequent mutations and reemergence of different strains of viruses, in order to avoid false negatives. It should also be noted that only the known and intended pathogens will be detected using such method and algorithm. Next-generation sequencing (NGS) technologies have emerged as powerful technique to perform massive parallel sequencing which enables us to sequence large number of samples with almost no limitation of pathogenic agents and targets on the genome. This ability not only provides the ability to diagnose atypical/rare pathogens but also provides molecular details on genotyping, virulence, drug resistance, and molecular epidemiology for the disease-causing virus.

5.2 Conventional Methods for Viral Diagnosis

Molecular Diagnostics Availability of whole genome sequences of viruses accelerated the development of quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) assays for diagnosis of viruses. Several public-funded research laboratories of different countries have developed quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) methods which are currently being used. All of these tests use hydrolysis probe chemistry and target-conserved regions of genes [5].

Immunodiagnosics Immunodiagnostic tests, especially immunochromatography-based rapid tests, are an attractive option as they provide quick point of care diagnosis of an infection. The basic principle for immune diagnosis or identification of the virus is based on the detection of viral antigen or antibodies made by the host against the pathogen. The most common antigenic targets for the detection of antigens consist of structural and nonstructural proteins; several antigen detection-based rapid kits for diagnosis of viruses have been developed [6].

5.2.1 Application of NGS in Viral Diagnosis

5.2.1.1 Viral Discovery Using Metagenomic Approach

Using metagenomic approach in a diagnostic procedure essentially means sequencing of all of the nucleic acid content and exploring the entire niche of the microbes. This not only limits the requirement of isolating and culturing the individual microbe but also enables identification of any pathogen without prior knowledge of its genomic sequences. In other words, it opens possibility to explore full diversity of virome/microbial community with novel and reemerging variants and coinfections [7, 8]. Selection of right sample type, time point, and type of nucleic acid (DNA/RNA) are important factors to be considered in viral metagenomics because secretion of expected type of virus may differ in different bodily secretions and may also differ at different time points of illness [9]. Furthermore, different sequencing platforms such as Illumina, Ion Torrent, PacBio, BGI, Nanopore, etc. have different read length and corresponding error rates [10]. It is well established that longer read is better classified in a taxonomic classification than the shorter ones. However, if the aim is classifying the microbe where exact genotyping and identification of variations are of lesser value, the error rates can be traded off with read lengths.

Further in the direction of data analysis, although we consider this approach as it doesn't require prior knowledge of the pathogen sequence, the metagenomic identification of any organism heavily depends upon existing pathogen sequence data (because classification is done on the basis of the similarity and the cluster where the pathogen belongs). This requirement limits the possibility of establishing the sensitivity and specificity of metagenomics-based viral pathogen detection approaches. Nevertheless, there are growing examples where this approach has been used in viral discovery of various animal and zoonotic viral diseases such as shaking mink syndrome astrovirus, Middle East respiratory syndrome (MERS), bat Influenza A virus, swine influenza C virus, bovine hepatitis virus, lateral-shaking inducing neurodegenerative agent (LINDA), etc. Various sequencing technologies, like Illumina, Ion Torrent, PacBio, and Nanopore, have provided different solutions for viral detection in clinical samples and achieved good diagnostic results.

5.2.1.2 Targeted Sequencing Approach

In contrast to metagenomic approach where full diversity of the microbes are identified without any biases, the targeted sequencing approach relies on enrichment of the target sequences which may be a partial genomic region highly conserved and

unique to the species or whole genome sequence. Targeted sequencing method is limited to samples where the viral genome size is small or viral load is low. In this method known panel of viruses can be targeted where each genome is amplified with overlapping amplicons. This approach has been successfully used to identify many viruses in the past, for example, HIV, influenza virus, HCV27, noroviruses, and rhabdoviruses. However, with the increase in size of genome and number of viruses in the panel, coverage of full genome with large number of primer sets becomes very difficult. For example, full genome amplification of Ebola virus required 11–19 set of primers. It becomes furthermore problematic if there are polymorphisms in the primer binding sites in viruses with high rate of mutations such as RNA viruses. In contrast to the sequencing of whole length of genome, amplicon-based detection of microbes has been much easier, cheap, and useful in the sense of diagnostic procedure. In amplicon-based detection of microbes, one or several conserved genomic regions are targeted which are unique to the species in question. Recently, the FDA has approved an Emergency Use Authorization for NGS-based diagnosis of COVID-19 provided by Illumina technology 32,641,848. The test consists of sequencing-based detection of 98 amplicons over 30 Kb of genome of SARS-CoV2. The strength of NGS-based viral diagnosis lies in large number of multiplexed sample (>3000 samples) with acceptable sensitivity and specificity. Previously, NGS-based detection of drug resistance-related mutations in HIV samples has been approved by the FDA, which displayed 95% sensitivity and specificity in detecting 342 HIV drug-resistant mutations.

Unlike targeted enrichment, hybridization-based enrichment methods are less tedious because it doesn't get affected with polymorphism and thus are suitable for sequencing of larger genomes. However, the hybridization-based enrichment is time-consuming and likely to miss genomic regions where probe differs more than 40% from target region.

5.3 Challenges and Limitations of NGS-Based Viral Diagnosis

Apart from the benefits mentioned, there are several challenges associated with utilization of NGS technology in viral diagnosis for clinical purposes.

- *Financial aspects:* The initial cost of the instrument and recurring costs for NGS-based diagnosis are high especially when the number of samples per run is not sufficient. Furthermore, storage and maintenance of large amount of data being generated, bioinformatic expertise for valid analysis, and public archival of clinical genomic data are challenging for a regular molecular diagnostic laboratory.
- *Incidental findings:* Although, because of the virtue of unbiased and exploratory nature of technology, NGS has emerged as very powerful tools for diagnosis, there are challenges of incidental findings. Incidental findings are those findings which are not part of the original or requested investigation but are of clinical importance to the patient or its family, for example, diagnosis of HIV in a viral

diagnosis protocol for other viruses using metagenomic approach. However, since the incidental finding is not requested by the patient, its reporting has been a matter of great ethical debate. Nevertheless, the NGS-based diagnosis based on target enrichment and PCR amplification has advantage because this only focuses on sequencing the genomic regions of pathogen of interest.

- *Regulatory issues:* In order to provide an authentic clinical diagnosis in a molecular diagnostic laboratory for any infectious or noninfectious disease, there are guidelines and accreditations provided by recognized agencies: standardization of protocols for sample processing, library preparation, sequencing, cutoff determination for minor variants, curation and updation of reference databases, and parameters of bioinformatic pipeline to rule out possible false negatives and false positives in the procedure.

References

1. Zaidi MB, Cedillo-Barron L, Gonzalez YAME, Garcia-Cordero J, Campos FD, Namorado-Tonix K et al (2020) Serological tests reveal significant cross-reactive human antibody responses to Zika and dengue viruses in the Mexican population. *Acta Trop* 201:105201
2. Mackay IM, Arden KE, Nitsche A (2002) Real-time PCR in virology. *Nucleic Acids Res* 30 (6):1292–1305
3. Goel S, Chakravarti A, Mantan M, Kumar S, Ashraf MA (2017) Diagnostic approach to viral acute encephalitis syndrome (AES) in Paediatric age group: a study from New Delhi. *J Clin Diagn Res* 11(9):DC25–DDC9
4. Malhotra B, Swamy MA, Janardhan Reddy PV, Gupta ML (2016) Viruses causing severe acute respiratory infections (SARI) in children ≤ 5 years of age at a tertiary care hospital in Rajasthan, India. *Indian J Med Res* 144(6):877–885
5. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V et al (2015) Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med* 7(1):99
6. Steingart KR, Henry M, Laal S, Hopewell PC, Ramsay A, Menzies D et al (2007) Commercial serological antibody detection tests for the diagnosis of pulmonary tuberculosis: a systematic review. *PLoS Med* 4(6):e202
7. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX et al (2016) Redefining the invertebrate RNA virosphere. *Nature* 540(7634):539–543
8. Fischer N, Indenbirken D, Meyer T, Lutgehetmann M, Lellek H, Spohn M et al (2015) Evaluation of unbiased next-generation sequencing of RNA (RNA-seq) as a diagnostic method in influenza virus-positive respiratory samples. *J Clin Microbiol* 53(7):2238–2250
9. Udugama B, Kadhiresan P, Kozlowski HN, Malekjahani A, Osborne M, Li VYC et al (2020) Diagnosing COVID-19: the disease and tools for detection. *ACS Nano* 14(4):3822–3835
10. Datta S, Budhauliya R, Das B, Chatterjee S (2015) Vanlalhmuaaka, veer V. next-generation sequencing in clinical virology: discovery of new viruses. *World J Virol* 4(3):265–276



Bioinformatic Application in COVID-19

6

Gurjot Kaur, Soham Mukherjee, and Shreya Jaiswal

Abstract

COVID-19 pandemic has seen massive application of in silico approaches to decipher the virus infectivity in humans as well as for the purpose of drug discovery. Data sharing has been a key to worldwide scientific collaboration and thus accelerating measures to counter the pandemic. Our chapter describes the main in silico approaches, their progression especially database generation and molecular modelling during the first year of the pandemic and emphasizes how these applications will contribute to pandemic-associated scientific discoveries, giving bioinformatics an important role for future tragedies.

Keywords

Bioinformatics · COVID-19

6.1 Introduction

COVID-19 pandemic started in December 2019 and was caused due to SARS-CoV-2 virus. Till the end of 2020, no specific treatment could be obtained. Drug and vaccine developments are progressing with the help of bioinformatic approaches. But first, let us delineate the historical use of in silico approaches in viral outbreaks.

G. Kaur (✉) · S. Jaiswal
School of Pharmaceutical Sciences, Shoolini University, Solan, India
e-mail: gurjotkaur@shooliniuniversity.com

S. Mukherjee
Faculty of Applied Sciences and Biotechnology, Shoolini University, Solan, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_6

6.1.1 Historical Facts in Previous Viral Outbreaks

Interestingly, COVID-19 is not the first viral disease where bioinformatic approaches were applied although the application of these approaches has magnified during the pandemic. The initial application was seen in Zika and Ebola virus as discussed below. Valuable lessons were learnt and amplified during the COVID-19 pandemic.

6.1.1.1 Zika Virus (2015–2016) and Ebola Virus (2014–2016) Outbreaks

Zika virus is caused by a flavivirus transmitted primarily by *Aedes* mosquitoes with mild symptoms lasting 2 to 7 days such as fever, rash, conjunctivitis, muscle and joint pain, and malaise or headache. Interestingly, Zika virus outbreak has been reported multiple times in the last century with the most recent one in 2015 in Brazil. As a countermeasure, Brazilian and American scientists came together for an open drug discovery collaborative effort. The scientific community recorded application of various computational strategies for drug repurposing and heavy usage of molecular docking methods on the viral proteins. These reports included protein homology modelling, X-ray crystallization structures, novel ligand, and protein discovery under the OpenZika project. However, due to a lack of corroborating in vitro and animal studies, many projects were shut down. On the other hand, the most recent Ebola virus outbreak took place in West Africa although it was first discovered in 1976 with fruit bats of the Pteropodidae family as natural hosts. It was previously called Ebola haemorrhagic fever, a severe, often fatal illness with a mortality rate of 50%. Drug discovery process for Ebola consisted of computational pharmacophore analysis of Ebola-active compounds, machine learning, in vitro testing, and generation of FDA-approved drugs (see review [1]). Table 6.1 summarizes the landmark computational studies for the two virus outbreaks.

6.2 COVID-19 Pandemic

By 2020, the COVID-19 patients became a major calamity as shown by COVID-19-positive cases and mortality on the WHO dashboard (<https://covid19.who.int/>). The WHO declared a state of global health emergency to coordinate scientific and medical efforts to rapidly develop a cure for patients [13]. The governments implemented social distancing and lockdowns to curb the spread. Many existing antiviral medications have been tried in hopes to slow or even cure the severely affected patients and thus decrease the mortality rate. One of the very promising strategies, therefore, was to use bioinformatic approaches as shown in Ebola and Zika outbreaks.

COVID-19 is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a new strain of coronaviruses that has been isolated from the Huanan Seafood Market, Wuhan, China, in December 2019. The identified primary reservoir is horseshoe bat, and transfer to human takes place through unknown intermediate hosts [14]. In general, the family of coronaviruses can cause respiratory,

Table 6.1 Computational approaches in Ebola and Zika drug discovery

Reference	Computational approach used in the study	Finding of the study
<i>Ebola virus</i>		
Brown, Lee [2]	<ul style="list-style-type: none"> • Structure-based screening (in silico) of millions of drug-like small molecules (approximately 5.4 million) targeting binding pocket of EBOV viral protein 35 (VP35), followed by in silico molecular docking of hits and NMR-based binding studies 	<ul style="list-style-type: none"> • Several compounds, especially pyrrolidinones, capable of binding to VP35 with high affinity and inhibiting polymerase co-factor activity identified
Ekins, Freundlich [3]	<ul style="list-style-type: none"> • Common features of pharmacophore analysis of four FDA-approved compounds exhibiting in vitro and in vivo activity against EBOV • Receptor-ligand pharmacophore analysis consisting of hydrophobic and hydrogen bonding features • In silico molecular docking of the FDA-approved drugs onto VP35 structure 	<ul style="list-style-type: none"> • Pharmacophore model for EBOV actives suggested common chemical features in the compounds and potentially similar target/mechanism • Receptor pharmacophore and molecular docking-supported VP35 could be a likely target for other FDA-approved drugs and analogues
Litterman, Lipinski [4]	<ul style="list-style-type: none"> • Computational validation of 55 small molecules with activity against EBOV and drug likeness property evaluation • Pan assay interference compounds (PAINS) computational filter to identify potentially problematic structures 	<ul style="list-style-type: none"> • Usage of medicinal chemistry, in silico and in vitro, and in vivo data could provide better measures for Ebola drug development
Veljkovic, Loiseau [5]	<ul style="list-style-type: none"> • Virtual screening was conducted of drugs against Ebola using EIIP/AQVN-based criterion 	<ul style="list-style-type: none"> • The selected (EIIP/AQVN) criterion was used as an efficient filter in screening for the inhibitors of Ebola virus infection. Drugs (approved and experimental) were selected by the criterion and represented the valuable source of therapeutics for the treatment of Ebola virus disease
Ekins, Freundlich [6]	<ul style="list-style-type: none"> • Bayesian machine learning computational models generated with molecules from EBOV replication assay and viral pseudotype entry assay. Models used to score drug libraries to identify potential inhibitors • Pharmacophore analysis conducted for best hits and further in vitro testing performed to check efficacy 	<ul style="list-style-type: none"> • Based on scoring by Bayesian models, three distinct molecules were identified which were previously not covered in literature pertaining to EBOV actives • Ligand pharmacophore analysis and further in vitro tests supported efficacy of the hits against EBOV
<i>Zika virus</i>		
Ekins, Perryman [7]	<ul style="list-style-type: none"> • Dengue virus crystal structure was used as template in homology model of ZIKV envelope protein. In silico screening performed through molecular docking of selected 	<ul style="list-style-type: none"> • Top 10 compounds from Prestwick Chemical Library including three antivirals (ritonavir, indinavir, and saquinavir) and a few antimalarials

(continued)

Table 6.1 (continued)

Reference	Computational approach used in the study	Finding of the study
	compounds from Prestwick Chemical Library to identify hits for in vitro testing	were identified as potential hits based on conformation scores
Ekins, Liebler [8]	<ul style="list-style-type: none"> • Homology models of ZIKV proteins generated through evolutionary relationships among flaviviruses. Best homology models screened according to GQME, QMEAN4 scoring, and Ramachandran plot analysis • Zika virion surface illustration generated by combining homology model with dengue virus envelope symmetry data 	<ul style="list-style-type: none"> • Homology models provided to serve as starting points for docking studies. Qualitative analysis of ZIKV virion compared to dengue virus cryo-EM virion outlined
Ekins, Liebler [8]	<ul style="list-style-type: none"> • Distributed computing on millions of devices to run docking of drugs (compounds) against structures (crystal) and homology models of Zika proteins 	<ul style="list-style-type: none"> • Homology modelling provided a way to accelerate new drug development in the absence of crystal structure of Zika proteins
Sahoo, Jena [9]	<ul style="list-style-type: none"> • Molecular docking analysis of four FDA-approved drugs (for dengue virus) and structure-based virtual screening of novel drug-like compounds against NS3 protein of ZIKV, which is essential for viral replication 	<ul style="list-style-type: none"> • Berberine was identified as the best FDA-approved antiviral with potential to be repurposed. Top 10 novel drug-like compounds sharing properties with berberine were identified, among which best hits were found to exhibit high binding affinity and interactions with key residues
Mottin, Braga [10]	<ul style="list-style-type: none"> • The study of molecular behaviour of helicase domain of NS3 (NS3h) in ZIKV in concert with ssRNA to elucidate NS3h activity and inhibition by molecular dynamics simulations 	<ul style="list-style-type: none"> • Molecular dynamic trajectories demonstrated presence of ssRNA and stabilize the RNA binding loop in NS3h so it maintains closed conformation. RNA binding loop conformation is also suggested to affect optimal ligand binding
Sharma, Murali [11]	<ul style="list-style-type: none"> • Induced fit docking of EGCG onto binding site of envelope protein to elucidate key interactions, molecular dynamic simulations, and drug-like property calculations 	<ul style="list-style-type: none"> • Interaction of EGCG with various residues was found, and it also provides protein conformation stabilization as analysed from molecular dynamics trajectories
Ramharack and Soliman [12]	<ul style="list-style-type: none"> • 'Per-residue energy decomposition pharmacophore' models for NS5 polymerase and NS5 methyltransferase using the molecules ribavirin and BG323. Virtual screening conducted using ZINC database and molecular docking performed for screened candidates 	<ul style="list-style-type: none"> • Virtual screening from ZINC database lead to identification of 23 candidates for NS5 polymerase and 18 candidates for methyltransferase. Compounds with the best docking scores were identified as potential actives (ZINC39563464 for NS5 polymerase and ZINC64717952 for NS5 methyltransferase)

VP35 viral protein 35, *NMR* nuclear magnetic resonance, *FDA* US Food and Drug Administration, *EBOV* Ebola virus, *EIIP* electron ion interaction potential, *AQVN* average quasi valence number, *EVD* Ebola virus disease, *ZIKV* Zika virus, *GQME* global model quality estimation, *QMEAN4* qualitative model energy analysis with linear combination of four statistical potential terms, *cryo-EM* cryogenic electron microscopy, *EGCG* epigallocatechin-3-gallate

gastrointestinal, hepatic, and central nervous system diseases. SARS-CoV-2 and its nearest neighbours in the phylogenetic tree, i.e. SARS-CoV and MERS-CoV, cause severe respiratory diseases. Its widespread transmission is due to travel and human to human contact [15]. SARS-CoV-2 is currently known to be sensitive to heat and UV rays and effectively destroyed with 75% ethanol, acetic acid, and chlorine-containing disinfectants [16].

6.2.1 Current Bioinformatic Efforts in COVID-19

We have highlighted the bioinformatic efforts that have been used in drug discovery research for COVID-19 till end of 2020.

6.2.1.1 Genomic Efforts: Sequencing Efforts

The whole genome sequences for the virus, SARS-CoV-2, have been isolated from patients from several countries including Brazil, China, Germany, and the USA. They were made publicly available as soon as they were sequenced to accelerate scientific research. There are currently more than 300 samples online. All the sequence samples were found to be closely related with few mutations pointing towards a common ancestor. For example, the Brazilian genome differed by three mutations to the Wuhan reference strain, and two of these three mutations were shared with a German sample. Efforts have been made to provide complete genome sequencing and thus aid the analysis of how the gene is translated to protein and what could be the protein functions especially in SARS-CoV-2 infectivity pathway. Genome sequencing is the starting point for all analysis regarding structure and function of resulting proteins. In addition, it provides knowledge about the origins of the SARS-CoV-2 virus and thus transmission profile. For the scientific research to rapidly move forward, it was crucial that the whole genome sequencing of SARS-CoV-2 is performed at a fast pace.

SARS-CoV-2 belongs to genus *Betacoronavirus* and subgenus *Sarbecovirus*. The virion or the infecting particle consists of an envelope containing a single positive-stranded RNA. The first genome, accession number NC_045512.2, was isolated from a patient in Wuhan, Hubei province, China, and named SARS-CoV-2 Wuhan-Hu-1. Current GenBank sequences and next-generation sequences stand at 39751 and 4266, respectively (data taken from NCBI-NLM SARS-CoV-2 Resources on November 12, 2020). Current estimates suggest a genome size of 29.9 kb and 11 open reading frames or ORFs. The organization of genes encoding the various proteins is shown below.

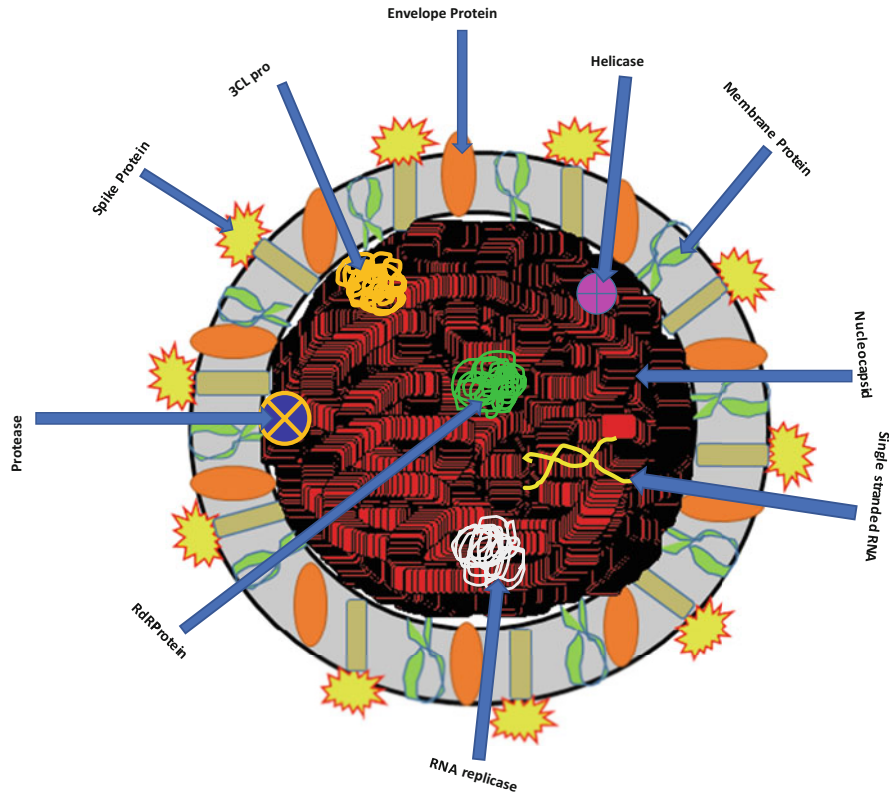


Fig. 6.1 Proteins on SARS-CoV-2 virion

5'-leader-UTR-replicase-ORF3-Spike (S)-Envelope (E)-Membrane (M)-Nucleocapsid (N)-3'UTR-poly (A) tail-3'-UTR end.

Figure 6.1 illustrates the cellular location of different proteins on SARS-CoV-2 virion. Unsurprisingly, the genome of SARS-CoV-2 is very similar to SARS-CoV (82%), bat-CoV-RaTG13 (96%), and bat-SL-CoVZC45 (86.9%). Main difference lies in the longer branch length to the bat viruses. Although mutations are being observed between various SARS-CoV-2 strains isolated from patients, the reported similarity is 99.98%. Phylogenetic analysis of 160 genomes has shown three main variants (classified as A, B, and C ancestral types) with certain mutations in specific variants, i.e. synonymous mutations T29095C and T8782C are identified in type A and type B, respectively, and non-synonymous mutations C28144T (Leu to Ser) and G26144T (Gly to Val) are detected in type B and type C, respectively [17]. This knowledge is being employed to identify genome-based community hotspots. For some mutations, radical changes in functionality of the protein, host specificity, or virus infectivity have been seen (Table 6.2). Mutational hotspots are located at positions 1397, 2891, 14408, 17746, 17857, 18060, 23403, and 28881. Mutational

Table 6.2 Common SARS-CoV-2 proteins and current structure models

Protein target	Function of the protein	Mutations	PDB ID of 3D structure
Mpro	Cysteine protease participates in the viral replication and cleaves the viral ORF1ab polyprotein at 11 sites	Mutation: R60C obtained from Mpro of Vietnam isolate revealed that the point mutation affects protease stability and binding of inhibitor [18]	6 LU7 6 W63 3M3V 6Y84 7BRO 6WQF 6 M03 6Y2E 6Y2F 6Y2G 6Y7M 6YB7 6LZE 6M0K 6YNQ 6YVF 6WNP 7BUY 7BRR 6WTT 7BRP 6YZ6 7BQY 6YT8
Spike glycoprotein (S)	The host-cell membrane is fused with the viral membrane. During the maturation of virus, Cleaving of spike protein is done to its subunits: virus is attached to the cell membrane through S1 subunit by interacting with host receptor. Fusion of the virus with human cell membranes is mediated by Ace2 and S2	Mutation: D614G, G476S, and V483A (most frequent) in RBD. They slow down the development of therapeutics Mutation: S19P and E329G may determine the host specificity of SARS-CoV-2. Important for molecular recognition, PCR testing kits, antiviral specificity, and vaccine development	6VYB 6VSB 6LXT 6LVN 2AJF 6VW1 6LZG 6M0J 6 W41 6YLA 6YM0 6WAQ 7BZ5 Complex of spike (6ACD) and ACE2 (2AJF) Complex of spike (6VXX) and ACE2 (1R42)

(continued)

Table 6.2 (continued)

Protein target	Function of the protein	Mutations	PDB ID of 3D structure
Polyprotein 1ab (ORF1ab-266–21,555 nucleic acids): NSP1–16 (ORF1ab)	<p>Replicase with multiple functions.</p> <p>Polyprotein consisting of 15 non-structural proteins, auto-proteolytically cleaved into multiple enzymes that form replicase-transcriptase machinery consisting of RNA-dependent RNA polymerase (RdRp), helicase, 3′–5′ exonuclease, endoRNase, and 2′-O-ribose methyltransferase.</p> <p>NSP15 encodes a nidoviral uridylyate-specific endoribonuclease that interacts with the NSP7/NSP8 complex.</p> <p>The antiviral activity of the STAT1 transcription factor is antagonized by ORF6 by sequestering IMPα/β1 on the rough ER/Golgi membrane. Human ubiquitin system with multiple members of the Cullin-2 E3 ligase complex interacts with the ORF10</p>	<p>Mutation: S24L in ORF8 protein strengthens the folding stability of the spike protein and ORF8 protein and shows female dominated pattern</p> <p>Mutation: A stabilizing mutation at position 321 in the endosome-associated-protein-like domain in NSP2 protein</p> <p>Mutation: A destabilizing mutation at position 192 in the NSP3 protein—useful for differentiating SARS-CoV-2 from SARS-related coronaviruses. Mutation: S723G and P1010I in the transmembrane helical segments of NSP2 and NSP3 determine the host specificity of SARS-CoV-2</p>	<p>NSP10–NSP16 complex: 6 W75 6YZ1</p> <p>NSP15: 6VWW 6 W01 6WXC 6WLC</p> <p>NSP16–NSP10 complex: 6W4H 6 W61 6WKS 6WQ3 6WRZ 6WVN 6WJT 6WKQ</p> <p>NSP3: 6VXS 6WEN 6 W02 6W6Y 6WCF 6WEY 6WOJ 6YWL 6YWK 6YWM</p> <p>NSP7 and NSP8 complex: 6WIQ 6WQD 6WTC 6YHU</p> <p>NSP9: 6W4B 6WXD 6W9Q</p>
RdRp or RNA-dependent RNA polymerase	It is required for the replication of viral genome and nucleic acid metabolism	Mutation: C14408T mutation is adjacent to the drugs targeting RdRp hydrophobic cleft	6 M71

(continued)

Table 6.2 (continued)

Protein target	Function of the protein	Mutations	PDB ID of 3D structure
Helicase	Required for viral genome replication and nucleic acid metabolism	Mutations: P504L and Y541C affected functional domain of NSP13 and even change the shape of ATP binding site in the helicase [19]	6XEZ
3'-5' exonuclease	Required for viral genome replication and nucleic acid metabolism	None yet reported	None yet available
EndoRNase (NSP15)	Required for viral genome replication and nucleic acid metabolism	None yet reported	7K1O 7K1L 6WXC 6 W01 6WLC 6X1B 6VWW
2'-O-ribose methyltransferase (encoded by non-structural NSP16)	Required for viral genome replication and nucleic acid metabolism. It binds N7-methyl guanosine cap and methylates the ribose 2'-O position of the first and second nucleotide of viral mRNA, which is essential to evade the immune system	Mutation: L111F, A116S, P236L, and two mutations localized towards termini, namely, P7S and N285D, have been identified in Indian isolates. These mutations are likely to alter protein dynamics, stability, and secondary structure and render pharmacological agents infective [20]	6YZ1 6W4H 6 W75 6 W61
Envelope protein (E)	This envelope protein self-assembles itself in the host membranes by forming ion channels and also acts as viroporin. It has a critical role in viral assembly and releases exerting viral pathogenicity	Mutation: Non-synonymous mutations in envelope protein observed in ~0.4% of SARS-CoV-2 whole genomes which might potentially affect propagation [21]	7K3G
Nucleocapsid protein (N)	This protein is responsible for the interaction of spike, envelope, and membrane proteins, which forms the nucleocapsid of the virus	Mutation: Several SARS-CoV-2 strains/substrains have exhibited mutations on serines 186, 197, and 202 which are phosphorylation sites involved in cell cycle control [22]	6M3M 6WZO 6WZQ 6WJI 6YUN 6ZCO 6WKP 7ACT 7ACS 2GIB 1SSK 7C22

(continued)

Table 6.2 (continued)

Protein target	Function of the protein	Mutations	PDB ID of 3D structure
			2CJR 6YI3 2JW8 2OFZ 2OG3 7CE0 7CDZ 6VYO

Original article references are given. Other references used for the table can be found in reviews: [23, 24]

variants are crucial to assess any possible drug resistance and COVID-19 clinical presentation. This information is also crucial for designing COVID-19 vaccines as well as rapid diagnostic assays. Structural genomic analysis has shown that the viral genome is composed of four structural proteins (spike-envelope-membrane-nucleo-capsid) and two non-structural proteins (main protease and RdRp) [23].

6.2.1.2 Protein Structure-Based Methods

Homology Modelling of SARS-CoV-2 Proteins

Considerable work has been accomplished in the development of homology models of SARS-CoV-2 proteins. Leading the way is the main viral protease or Mpro. Other proteins that have been heavily modelled through homology modelling are spike (S) protein, ACE2 human target, and RdRp. Table 6.2 provides the current list of protein targets and highlights whether a specific target is over- or underutilized for the computational studies.

The knowledge of three-dimensional structure of proteins is crucial to develop drugs that can modulate the protein's action. The structural information, obtained through X-ray crystallography or cryogenic electron microscope or nuclear magnetic resonance, provides information on binding sites and the mechanism of action/inhibition. The three-dimensional structures are also crucial to molecular docking studies as they serve as starting point. These models have been used for docking to understand mechanisms of viral infection and possible treatments. Currently, there are over 400 (total) structures available for various SARS-CoV-2 targets on RCSB PDB.

Molecular Docking Studies

Docking simulations have been performed for treatment through both small molecules and antibodies. Small molecule strategies require a target protein structure and screening of molecules that bind this protein using molecular docking and validation using molecular dynamic simulations. Many studies have been performed

in a very short span of time with the rapidly made available targets (Table 6.2). In the second method, antigen-antibody docking simulation has predicted high-affinity binding of human antibodies to SARS-CoV-2 proteins. Human antibody CR3022 is shown to possess a high affinity for spike protein [24].

Unfortunately, many early studies lacked proper validation of molecular docking data (redocking or molecular dynamics), and thus reproducibility of the results is highly doubtful. This trend was seen for many drug or herbal chemical candidates in the early days of COVID-19. Studies that came a little later validated the data by molecular dynamics and provided important parameters regarding drug/herbal chemical stability inside the binding pocket of protein target and possible hydrogen bonding and hydrophobic interactions during the simulations [25–28]. Very few direct validations of molecular docking protocol by redocking with the known ligand, i.e. native ligand on the X-ray crystal structure, were performed in the early days [1, 24, 29].

Drug Repurposing

As drug discovery is a long process and initiating it from a completely new drug candidate may result in a long wait while the COVID-19 pandemic was gaining momentum, drug repurposing seemed to be a much rapid alternative, previously employed for many related (MERS and SARS-CoV) and unrelated diseases. In this process, many drugs have been screened [1, 24]. These candidate molecules may be used in SARS-CoV-2 enzyme assays, antagonism of protein mechanism, or decelerating viral infectivity. It is hypothesized that drugs previously used in other viral diseases could also be used. An important aspect would be to make sure that the drugs are, in general, available for further experimentation and development in case they show promise. It will be futile to develop drugs that are only available as structural moieties or are at an investigational stage only. A simple strategy to drug repurposing that shows promise is screening a class of compounds rather than random drug screening. FDA-approved HIV-1 protease inhibitors, e.g. atazanavir and ritonavir, and hepatitis C NS3/4A protease inhibitors, e.g. lopinavir, have been successfully docked into the Mpro active site of SARS-CoV-2. Currently, China and South Korea have treated COVID-19 patients with Kaletra, the combination of lopinavir 200 mg/ritonavir 50 mg with some benefits [30].

6.2.1.3 Database Generation

It is needless to say that databases are very important to accelerate research in the current times especially because they provide a reference point to scientists worldwide. Virtual databases provide the advantage of large storage capacity while being continuously updated. In SARS-CoV-2, this has provided a better understanding of the virus's origin after extensive comparisons between genomic data online. Genomic comparisons also helped in producing specific primers for RT-PCR detection kits as early as possible. Many databases have proven extremely useful in the pursuit of treatment strategies for COVID-19. While primary databases such as nucleotide sequence databases and protein sequence databases are extensively used to submit as well as obtain sequencing information for COVID-19 targets, secondary databases

Table 6.3 COVID-19-specific online databases

Database	Information provided by the database	References
2019nCoV-VR	For studies on classification of virus (viral taxonomy), evolution (changes) of genome, molecular diagnosis, and development of drugs	[31]
The Genome Detective Coronavirus Typing Tool	Used for the tracing of new viral mutations in SARS-CoV-2 genome sequences accurately	[32]
CoronaVR	It is used for the identification of potential epitope-based (B cells and T cells) vaccine candidates, siRNA-based (subsection of therapeutic section) therapeutic regimens, and diagnostic primers, that can be used in testing kits to identify COVID-19 virus	[33]
CoV-AbDab	It contains a relevant set of data which includes cross-neutralization antibody/nanobody origin evidences, full variable domain sequence, assignments of germline, regions of epitope, links to relevant PDB entries, comparative modelling (homology models), and literature. It is also used for the aid in preceding exposure diagnosis or to assist in predicting the efficacy of vaccine. It is used for the understanding of antibody at molecular level, which is able to attach with a <i>Betacoronavirus</i> antigen that could be relevant in treating COVID-19 (SARS-CoV-2) infection	[34]

such as Prosit, PRINTS, Pfam, InterPro, PhenomicDB, or a genotype-phenotype database provide support to understand protein structure and functionality, while molecular structure databases such as Protein Data Bank, SCOP, CATH, and PubChem are important to obtain structural information of molecular targets. A few COVID-19-specific databases are listed in Table 6.3.

Due to the widespread effect of the COVID-19 pandemic across countries and to allow easy sharing of scientific information between scientists, these databases have come up. While the WHO COVID-19 database provides general information on sociodemographics of the mortality and infected individuals, ViPR or Virus Pathogen Database and Analysis Resource has been updated to include information on SARS-CoV-2 and contains SARS-CoV-2-related data, tools, and analysis. CoV-AbDab provides data on COVID-19 antibodies. The International Nucleotide Sequence Database Collaboration (INSDC) has released a public statement entitled ‘INSDC Statement on SARS-CoV-2 sequence data sharing during COVID-19’, which highlights the importance of sharing SARS-CoV-2 sequence data within the international scientific community. The INSDC recommends that all researchers working with SARS-CoV-2 sequence data submit both their raw and consensus—or assembled—SARS-CoV-2 data to the INSDC databases, which are freely available to the scientific community. COVID-19 data portal EMBL-EBI (<https://www.covid19dataportal.org/>) enables data sharing throughout the globe. The initiative facilitates international collaboration to accelerate scientific discovery, monitor the

pandemic, and help develop treatments and a vaccine for the new coronavirus. Other SARS-CoV-2 resources can be accessed by NCBI at <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

6.2.1.4 Other Approaches

It should be noted that previous virus research did not employ many of the above discussed tools. Ebola and Zika viral research were the first application of bioinformatic research in the viral disease field, and COVID-19 marks the introduction of drug discovery by using *in silico* approaches as constructive options. All the current bioinformatic tools have helped scientist in accelerating drug/vaccine development for SARS-CoV-2. Since the advance of bioinformatics in 1990, many tools exist that are used in addition to genome sequencing and molecular modelling. Softwares (both online and offline) are being extensively utilized for sequence analysis, complete genome sequencing, expressed sequence tags, identification of unknown genes, discovery of splice variants, causes of differences between viruses, pharmacogenetics, next-generation sequencing, etc. Multiple alignment tools to perform alignment of experimentally obtained genetic sequences are important for sequence comparisons and sequence-based database searches and additionally help in phylogenetic analysis. This is evident from the use of RNA sequencing from broncho-alveolar fluid samples of SARS-CoV-2 patients to identify its origins. The phylogenetic analysis revealed almost 90% similarity of virus sequences to betacoronaviruses from bat [35]. BLAST or Basic Local Alignment Search Tool has also been widely used to compare sequences whether nucleic acid or amino acid. BLAST uses the already available information on biological sequences from organisms to understand the genetic relationship with other species. For example, SARS-CoV-2 genome sequence similarity with viral metagenomes in pangolin has been seen. In particular, the availability of these bioinformatic approaches was very important in the discovery of newer drugs for COVID-19 as an understanding of genetic sequence needed to follow up by the protein analysis, i.e. function of the gene. This was made possible by comparing related sequences and thus similar functionality. Once the protein function could be determined (coupled to experimental evidence), the already known drugs against related targets could be taken for drug repurposing. In addition, drugs could be developed using computational approaches once this basic understanding is obtained [29].

In addition, due to the rapid research in the field of COVID-19, a comprehensive repository of knowledge about SARS-CoV-2, its proteins, mechanism of infection, and more has been created online. Vaccine design is also utilizing computational methods to design multiepitopes against SARS-CoV-2. The vaccine design includes prediction of potential epitopes from antigenic protein sequences and construction of vaccine, followed by molecular docking simulation to assess the binding affinity to the protein. For example, antigenic epitopes from spike glycoprotein, nucleocapsid, ORF3a, and non-structural proteins have been attempted (reviewed in [24]). Vaccine design has benefitted immensely by the constructed structural genomic and interactomic road maps that describe viral infection molecular mechanisms. An example is the X-ray crystal structure of RBD in complex with human antibody

CR3022 (6 W41) for epitope engineering. T- and B-cell epitopes have been identified too through informatics (details in [23]). Development of SARS-CoV models using ECFP6 descriptors and the Bayesian algorithm has been attempted to develop an assay control software (Ekins 2020). The method is fast, does not require crystal structures, and enables scoring of small-molecule structures against many models simultaneously. System pharmacology approaches have also provided important insights into promising antiviral drugs against SARS-CoV-2 based on pathogenesis mechanism and host specificity. Novel algorithms are being used in COVID-19 research. These are mainly network-based algorithms and expression-based algorithms. Network-based algorithms generate networks, for example, drug target network or human protein interactome, and using these networks, putative drug candidates are identified. One such study generated HCoV-host interactome and integrated various drugs specific for targets [36]. Some expression-based algorithms have also aided drug repurposing by linking ACE2 to SARS-CoV-2 in the early days and thus providing two potential repurposed drugs that change ACE2 expression. Functional analysis of genomes is done in parallel to identify the cellular functions of gene products coupled to transcriptomics, proteomics, metabolomics, phenomics, and even systems biology. See reviews [23, 24] for details. Table 6.4 provides a few examples of applications of all the above-mentioned approaches in COVID-19.

6.3 Conclusion

There are many lessons that can be learnt from the application of *in silico* approaches in the viral pandemic of COVID-19. On the positive side, molecular modelling provides the advantage of reduced cost for faster development of a drug candidate. This is propelled by databases and collaborative efforts of scientists. Scientists can thus directly compare the various drug candidates in these databases and assess for further development.

This was the third instance of application of these techniques, and therefore, important lesson would be to rely more on high-resolution crystal structures rather than homology models. Otherwise, it will be very difficult to process the huge repository of docking results generated every time. Also, database generation should become a priority to increase the collaborative capacity and, therefore, data validation by scientists throughout the world. In addition, if the molecular docking is performed on commercially available drugs and herbal chemicals, it would help us limit the number of docked molecules yet keeping the feasibility of taking the drug candidate to the next stage of drug development. Molecular docking of chemicals that are not readily available leads to extra steps of extraction or synthesis that could be avoided. It is thus understandable that a collaborative approach where scientists team up to work on different aspects such as bioinformatics, chemical synthesis, *in vitro* testing, and *in vivo* testing will be fruitful. Interestingly, the comprehensive information about virus pathogenesis in the human body is still not available and thus hinders progress in drug development.

Table 6.4 Examples of bioinformatic approaches used in COVID-19

Bioinformatic approach used in COVID-19 research	Reference examples and details
Whole genome sequencing	[37]—RT-PCR, next-generation sequencing on specimens coupled to characterization at molecular level, phylogenetic analysis, and B- and T-cell epitope prediction for the sequences of Indian SARS-CoV-2. Complete (29,851 nucleotides) genomes with 99.98% identity with Wuhan seafood market pneumonia virus (accession number: NC 045512) were obtained. In the receptor binding domain, prediction of linear B-cell epitopes of spike protein of S1 domain and a conformational epitope was identified. Results confirmed two different introductions into the country
Sequence comparisons	[18]—The SARS-CoV-2 genome reported from 13 different countries was investigated to identify mutations. High identity (>99%) was seen amongst the 13 genomes. By sequence comparison, country-specific distinctive mutations in the vital proteins of SARS-CoV-2 were identified (vital proteins, i.e. replicase polyprotein, spike glycoprotein, envelope protein, and nucleocapsid protein). Mutational effects on function were investigated using various in silico approaches
Algorithms	[38]—Development of simple algorithm to help the early detection in the patients who are infected by SARS-CoV-2. Identification of individual predictors of COVID-19 development and establishment of prediction model with a learning set of 322 COVID-19 patients (training datasheet/set) and a set of 317 COVID-19 patients were validated. The analysis identified age, lactate dehydrogenase, and CD4 count as good predictors of COVID-19 progression giving an algorithm of (age × LDH)/CD4 count
Machine learning	[39]—These models were built for the prediction of protein-protein interactions between the viral proteins and proteins present in humans. A total of 1326 potential human target proteins of SARS-CoV-2 were predicted by the proposed ensemble model and validated using gene ontology and KEGG pathway enrichment analysis
Systems biology	[40]—166 modified herbal Chinese formulae and 1212 phytoconstituents that were containing b-sitosterol, stigmasterol, and quercetin were chosen. Using complex system entropy and unsupervised hierarchical clustering, 18 herbal formulae showed promise as COVID-19 candidates. 12 clusters of molecules yielded 8 pharmacophore families of structures following scaffold analysis, self-organizing mapping, and cluster analysis
Drug repurposing	[41]—The amalgamation of drugs, viz. pirfenidone and melatonin, was identified for COVID-19 by using system biology and artificial intelligence-based approaches. GUILDify v2.0 web server was also used to confirm the results

(continued)

Table 6.4 (continued)

Bioinformatic approach used in COVID-19 research	Reference examples and details
Vaccine development	[42]—Characterization of spike glycoprotein to obtain immunogenic epitopes using immunoinformatic approach. 13 major histocompatibility complex-(MHC) I and 3 MHC-II epitopes having antigenic properties were chosen as they were linked to specific linkers to build vaccine components and molecularly docked on Toll-like receptor-5 to get binding affinity

In conclusion, due to the heavy inundation with incomplete or unvalidated molecular modelling studies, it is warranted that proper steps are taken to manage pseudoscience propagation.

6.4 Future Outlook

Application of bioinformatic approaches in Ebola, Zika, and COVID-19 pandemic has shown that these approaches can be used for planning and developing antiviral drugs and support pandemic situations when properly streamlined. Target identification, interaction mapping, understanding structure activity relationships, and molecular docking and dynamics could provide important tools for elucidating underlying pathogenesis and its targeting by novel drug candidates. When coupled to proper experimental testing, the drug development will have far-reaching results. We are able to increase the safety of the drug molecules, i.e. by understanding the physical-chemical properties as well as probability of success. Time must be spent on elucidating underlying mechanisms so as to provide relief to clinical symptoms of COVID-19.

Acknowledgements GK conceptualized and wrote the chapter. SM and SJ prepared the Tables and Fig. 6.1. This work is part of the project MCB200120 funded through COVID-19 HPC Consortium.

References

1. Ekins S, Mottin M, Ramos P, Sousa BKP, Neves BJ, Foil DH et al (2020) Deja vu: stimulating open drug discovery for SARS-CoV-2. *Drug Discov Today* 25(5):928–941
2. Brown CS, Lee MS, Leung DW, Wang T, Xu W, Luthra P et al (2014) In silico derived small molecules bind the filovirus VP35 protein and inhibit its polymerase cofactor activity. *J Mol Biol* 426(10):2045–2058
3. Ekins S, Freundlich JS, Coffee M (2014) A common feature pharmacophore for FDA-approved drugs inhibiting the Ebola virus. *F1000Res* 3:277
4. Litterman N, Lipinski C, Ekins S (2015) Small molecules with antiviral activity against the Ebola virus. *F1000Res* 4:38

5. Veljkovic V, Loiseau PM, Figadere B, Glisic S, Veljkovic N, Perovic VR et al (2015) Virtual screen for repurposing approved and experimental drugs for candidate inhibitors of EBOLA virus infection. *F1000Res* 4:34
6. Ekins S, Freundlich JS, Clark AM, Anantpadma M, Davey RA, Madrid P (2015) Machine learning models identify molecules active against the Ebola virus in vitro. *F1000Res* 4:1091
7. Ekins S, Perryman AL, Horta AC (2016) OpenZika: an IBM World Community Grid project to accelerate Zika virus drug discovery. *PLoS Negl Trop Dis* 10(10):e0005023
8. Ekins S, Liebler J, Neves BJ, Lewis WG, Coffee M, Bienstock R et al (2016) Illustrating and homology modeling the proteins of the Zika virus. *F1000Res* 5:275
9. Sahoo M, Jena L, Daf S, Kumar S (2016) Virtual screening for potential inhibitors of NS3 protein of Zika virus. *Genomics Inform* 14(3):104–111
10. Mottin M, Braga RC, da Silva RA, Silva J, Perryman AL, Ekins S et al (2017) Molecular dynamics simulations of Zika virus NS3 helicase: insights into RNA binding site activity. *Biochem Biophys Res Commun* 492(4):643–651
11. Sharma N, Murali A, Singh SK, Giri R (2017) Epigallocatechin gallate, an active green tea compound inhibits the Zika virus entry into host cells via binding the envelope protein. *Int J Biol Macromol* 104(Pt A):1046–1054
12. Ramharack P, Soliman MES (2018) Zika virus NS5 protein potential inhibitors: an enhanced in silico approach in drug discovery. *J Biomol Struct Dyn* 36(5):1118–1133
13. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A et al (2020) World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). *Int J Surg* 76:71–76
14. Guo YR, Cao QD, Hong ZS, Tan YY, Chen SD, Jin HJ et al (2020) The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status. *Mil Med Res* 7(1):11
15. Ralph R, Lew J, Zeng T, Francis M, Xue B, Roux M et al (2020) 2019-nCoV (Wuhan virus), a novel coronavirus: human-to-human transmission, travel-related cases, and vaccine readiness. *J Infect Dev Ctries* 14(1):3–17
16. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W et al (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579(7798):270–273
17. Forster P, Forster L, Renfrew C, Forster M (2020) Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* 117(17):9241–9243
18. Khan MI, Khan ZA, Baig MH, Ahmad I, Farouk AE, Song YG et al (2020) Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: an in silico insight. *PLoS One* 15(9): e0238344
19. Ugurel OM, Mutlu O, Sariyer E, Kocer S, Ugurel E, Inci TG et al (2020) Evaluation of the potency of FDA-approved drugs on wild type and mutant SARS-CoV-2 helicase (Nsp13). *Int J Biol Macromol* 163:1687–1696
20. Azad GK (2020) Identification of novel mutations in the methyltransferase complex (Nsp10-Nsp16) of SARS-CoV-2. *Biochem Biophys Rep* 24:100833
21. Hassan SS, Choudhury PP, Roy B (2020) SARS-CoV2 envelope protein: non-synonymous mutations and its consequences. *Genomics*
22. Tung HYL, Limtung P (2020) Mutations in the phosphorylation sites of SARS-CoV-2 encoded nucleocapsid protein and structure model of sequestration by protein 14-3-3. *Biochem Biophys Res Commun* 532(1):134–138
23. Chellapandi P, Saranya S (2020) Genomics insights of SARS-CoV-2 (COVID-19) into target-based drug discovery. *Med Chem Res*:1–15
24. Wang X, Guan Y (2020) COVID-19 drug repurposing: a review of computational screening methods, clinical trials, and protein interaction assays. *Med Res Rev*
25. Rolta R, Yadav R, Salaria D, Trivedi S, Imran M, Sourirajan A et al (2020) In silico screening of hundred phytocompounds of ten medicinal plants as potential inhibitors of nucleocapsid

- phosphoprotein of COVID-19: an approach to prevent virus assembly. *J Biomol Struct Dyn*:1–18
26. Costa AN, de Sa ERA, Bezerra RDS, Souza JL, Lima F (2020) Constituents of buriti oil (*Mauritia flexuosa* L.) like inhibitors of the SARS-Coronavirus main peptidase: an investigation by docking and molecular dynamics. *J Biomol Struct Dyn*, 1–8
 27. Krupanidhi S, Abraham Peele K, Venkateswarulu TC, Ayyagari VS, Nazneen Bobby M, John Babu D et al (2020) Screening of phytochemical compounds of *Tinospora cordifolia* for their inhibitory activity on SARS-CoV-2: an in silico study. *J Biomol Struct Dyn*:1–5
 28. Ahmad S, Abbasi HW, Shahid S, Gul S, Abbasi SW (2020) Molecular docking, simulation and MM-PBSA studies of *nigella sativa* compounds: a computational quest to identify potential natural antiviral for COVID-19 treatment. *J Biomol Struct Dyn*:1–9
 29. Villas-Boas GR, Rescia VC, Paes MM, Lavorato SN, Magalhaes-Filho MF, Cunha MS et al (2020) The NEW Coronavirus (SARS-CoV-2): a comprehensive review on immunity and the application of bioinformatics and molecular modeling to the discovery of potential anti-SARS-CoV-2 agents. *Molecules* 25(18):4086
 30. Lim J, Jeon S, Shin HY, Kim MJ, Seong YM, Lee WJ et al (2020) Case of the index patient who caused tertiary transmission of COVID-19 infection in Korea: the application of Lopinavir/ritonavir for the treatment of COVID-19 infected pneumonia monitored by quantitative RT-PCR. *J Korean Med Sci* 35(6):e79
 31. Zhao WM, Song SH, Chen ML, Zou D, Ma LN, Ma YK et al (2020) The 2019 novel coronavirus resource. *Yi Chuan* 42(2):212–221
 32. Cleemput S, Dumon W, Fonseca V, Abdool Karim W, Giovanetti M, Alcantara LC et al (2020) Genome detective coronavirus typing tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics* 36(11):3552–3555
 33. Gupta AK, Khan MS, Choudhury S, Mukhopadhyay A, Sakshi, Rastogi A et al (2020) CoronaVR: a computational resource and analysis of epitopes and therapeutics for severe acute respiratory syndrome coronavirus-2. *Front Microbiol* 11:1858
 34. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM (2020) CoV-AbDab: the coronavirus antibody database. *Bioinformatics*
 35. Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y et al (2020) Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm Sin B* 10(5):766–788
 36. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F (2020) Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* 6:14
 37. Yadav PD, Potdar VA, Choudhary ML, Nyayanit DA, Agrawal M, Jadhav SM et al (2020) Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian J Med Res* 151(2 & 3):200–209
 38. Li Q, Zhang J, Ling Y, Li W, Zhang X, Lu H et al (2020) A simple algorithm helps early identification of SARS-CoV-2 infection patients with severe progression tendency. *Infection* 48(4):577–584
 39. Dey L, Chakraborty S, Mukhopadhyay A (2020) Machine learning techniques for sequence-based prediction of viral-host interactions between SARS-CoV-2 and human proteins. *Biom J*
 40. Luo L, Jiang J, Wang C, Fitzgerald M, Hu W, Zhou Y et al (2020) Analysis on herbal medicines utilized for treatment of COVID-19. *Acta Pharm Sin B* 10(7):1192–1204
 41. Artigas L, Coma M, Matos-Filipe P, Aguirre-Plans J, Farres J, Valls R et al (2020) In-silico drug repurposing study predicts the combination of pirfenidone and melatonin as a promising candidate therapy to reduce SARS-CoV-2 infection progression and respiratory distress caused by cytokine storm. *PLoS One* 15(10):e0240149
 42. Bhattacharya M, Sharma AR, Patra P, Ghosh P, Sharma G, Patra BC et al (2020) Development of epitope-based peptide vaccine against novel coronavirus 2019 (SARS-CoV-2): Immunoinformatics approach. *J Med Virol* 92(6):618–631

Part II

Infection Transcriptomics: Transcriptomics Applications for Human Pathogens



Emerging Transcriptomic Approaches to Decipher Mycobacterial Complexities

7

Jasmine Samal, Nilofer Naqvi, Yashika Ahuja, Neha Quadir, P. Manjunath, Faraz Ahmad, Mohd. Shariq, Anwar Alam, Avantika Maurya, and Nasreen Z. Ehtesham

Abstract

The word “transcriptome” generally refers to overall RNA species transcribed from DNA (or RNA itself) or transcripts that are present in a cell. The transcriptome includes messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and even noncoding RNAs, such as miRNA, having a regulatory function. TB pathogenesis is highly dynamic and is determined by the interaction between the host defense strategy and the tactics employed by *Mycobacterium* species to survive inside the host. Recently, several reports suggested that differential expression of host miRNAs serves as hallmark of disease progression at both cellular and organism level in various diseases, including tuberculosis. Next-generation sequencing (NGS) is a revolutionized massive parallel sequencing technique that uses RNA-seq, which is a recent NGS-based technology for transcriptome profiling which is most often used to analyze differential gene expression and differential splicing of mRNA. RNA-seq is more sensitive than other technologies such as microarray as it directly determines the cDNA and thus provides a great insight about physiological state of cell, healthy or diseased. In this chapter, we present a detailed overview of RNA-seq technology and its data analysis, applications, and advances made so far. With the advancement in NGS, RNA-seq has also developed. Single RNA-seq is the further refinement of RNA-seq analysis in a single cell, which allows studying the complex biological processes especially useful in cancer studies. Until recently, the technique was limited to mRNA expression of either pathogen or host cell. However, in the recent years, more advanced RNA-seq technology enables “dual

J. Samal · N. Naqvi · Y. Ahuja · N. Quadir · P. Manjunath · F. Ahmad · M. Shariq · A. Alam · A. Maurya · N. Z. Ehtesham (✉)
Inflammation Biology and Cell Signalling Laboratory, ICMR National Institute of Pathology, Safdarjung Hospital Campus, New Delhi, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

107

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_7

RNA-seq” analysis for better understanding of host-pathogen interaction. Dual RNA-seq does not require pre-designed species-specific probes and thus provides a better understanding of the interaction, virulence factor, or immune response mechanism. With increasing sensitivity of dual RNA-seq, it is now being applied to understand the host-pathogen interaction in tuberculosis, which is the key study in development of therapeutic strategies to control “*Mycobacterium tuberculosis*.” With more refinement at lower cost, RNA-seq is expected to replace the microarray technique in the near future.

Keywords

RNA · miRNA · NGS · M.tb · Diagnostics

7.1 Introduction

The word “transcriptome” generally refers to overall RNA species transcribed from DNA (or RNA itself) or transcripts that are present in a cell. The transcriptome includes messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and even noncoding RNAs, such as miRNA, having a regulatory function. Recently, there emerged a high-throughput sequencing technology which involves analysis of RNA at single nucleotide level, called RNA sequencing (RNA-seq). RNA-seq is an NGS-based sequencing technology with many applications, such as mRNA expression quantification, differential splicing of mRNA, fusion genes detection, and RNA editing. RNA-seq is a powerful tool that directly determines the cDNA and hence provides a great insight of genomic functions and physiological state of cell. RNA-seq has advantage over qPCR and microarrays as it provides complete transcriptome of the organism including coding and noncoding genes, intergenic regions, and small RNAs [1]. *Mycobacterium tuberculosis* (*M. tb*) is the causative agent of tuberculosis and is the leading cause of death globally [2]. It is estimated that around one-fourth of the world population is infected with TB; however, only 5%–10% of these cases advance to active TB in their lifetime [2]. In this chapter, we will discuss emerging transcriptomic approaches to decipher complexities related to *Mycobacterium tuberculosis*.

7.2 MicroRNAs: Biogenesis and Mechanism of Action

MicroRNAs (miRNAs) are a class of highly conserved, small, noncoding single-strand RNA molecules that are approximately 18–25 nucleotides in length. They play an important role in the posttranslational expression of genes and are involved in cell development, differentiation, proliferation, apoptosis, etc. [3]. miRNAs are divided into two classes, the intragenic and intergenic miRNAs, based on their genomic loci. Intergenic miRNAs are located between two genes and can be monocistronic (with their own promoters) or polycistronic, wherein various

miRNAs are transcribed together, while the intragenic miRNAs are known to be present in the introns or exons of genes and are co-regulated with their host genes by Pol II. miRNA biogenesis pathway begins with the transcription of miRNA by RNA Pol II to form hairpin-like structure called the pri-miRNA. These are then cleaved by a nuclease called Drosha, resulting in the formation of pre-miRNA. The pre-miRNA is transported to the cytoplasm by exportin-5 through a Ran-GTP-dependent mechanism and is further excised by Dicer to form the mature miRNA duplex. A single strand of this miRNA duplex then binds to the other accessory proteins such as Ago-2 and GW182 and is incorporated in the RNA-induced silencing complex (RISC) for posttranscriptional regulation of mRNAs [4]. The microRNAs usually bind to a specific region, about 5–8 nucleotides long, in the 3'-UTR region of the messenger RNA, resulting in the degradation of the target mRNA or inhibition of translation.

7.2.1 miRNAs in Tuberculosis: New approaches for Diagnosis and Host-Directed Therapy

TB pathogenesis is highly dynamic and is determined by the interaction between the host defense strategy and the tactics employed by *Mycobacterium* species to survive inside the host. Currently, TB diagnostic methods primarily focus on diagnosing active tuberculosis, thus demanding more definitive biomarkers that differentiate latent and active TB for a more specific and efficient diagnosis. Owing to the limitations in traditional methods for diagnosis of TB, researchers have emphasized the need for development of diagnostic tools based on host biomarkers that can be used for assessment of disease status and monitoring of treatment outcomes. In pursuit of the same, several classes of molecules are currently being scrutinized as prospective biomarkers, miRNAs being one of these biomarkers.

7.2.2 Role of Host miRNAs in Tuberculosis

Recently, several reports suggested that differential expression of host miRNAs serves as hallmark of disease progression at both cellular and organism level in various diseases, including tuberculosis. Interestingly, the pathogen has evolved various mechanisms to survive inside the macrophages and establish dormancy/latency. Notably, miRNAs have been emerging as critical regulators of immune response in *M. tb* infection. Here, we mention some of the important host-bacterial interactions that are modulated by miRNAs in tuberculosis.

7.2.2.1 Host miRNAs Regulate Autophagy and Phagolysosome Maturation in Tuberculosis

Autophagy is a well-known cell-autonomous defense mechanism conducted against many intracellular microorganisms. To ensure their own survival, pathogenic bacteria modulate many host cell processes including autophagy. Accumulating

evidences suggest that host miRNAs regulate several autophagy-related genes (ATGs), thus favoring mycobacterial survival. Wang et al. [5] showed that the expression of host miR-155 was enhanced after mycobacterial infection using in vitro and in vivo models. This enhanced expression accelerated the formation of phagolysosome and helps in eradicating the infection. Kim et al. [6] showed that miR-125a inhibits autophagy and antimicrobial effects against *Mycobacterium tuberculosis* by targeting the UV radiation resistance-associated gene (UVRAG). miR-20a inhibits autophagy by targeting ATG7 and ATG16L1, hence favoring mycobacterial survival inside the host macrophages. miR-125a-3p, miR-33, miR-144-3p, miR-23a-5p, and miR-142-3p are potential inhibitors of autophagy in *Mycobacterium tuberculosis* infection. Notably miR-155, miR-17-5p, and miR-26a target Ras homolog enriched in the brain (Rheb), Mcl-1, STAT3, and KLF4, respectively, play a role by inhibiting the infection.

One of the well-studied mechanisms of immune escape by *M. tb* is mediated by arresting phagosome-lysosome maturation. Prevention of acidification of phagosome with lysosomal enzymes, such as cathepsin proteases, plays a crucial role in persistence of *M. tb* inside host macrophages. miR-106-5p is known to target cathepsin S and inhibits lysosomal enzymatic activity. Another microRNA, miR-142-3p, inhibits internalization by phagosomes by targeting N-WASP and PKC-alpha genes [7]. Polarization of macrophages as M2 renders macrophages as anti-inflammatory and poorly microbicidal. A transcription factor, KL4, drives M2 polarization of macrophages and is negatively regulated by miR-26a. Notably, in patients with TB, miR-26a is found to be downregulated [8].

7.2.2.2 Host miRNAs Regulate Innate Immunity in Tuberculosis

Innate immune cell activation is regulated by miR-155, miR-146a, miR-21, and miR-9. miR-125a-3p can inhibit antimicrobial responses and host defense against *M. tb* infection by targeting the gene encoding autophagy UV radiation resistance-associated protein [6]. miR-125b inhibits the production of TNF- α in alveolar macrophages during *M. tb* infection. TNF- α is a well-known pro-inflammatory cytokine for the *M. tb* clearance. Targeting miR-125a and miR-125b will be helpful in increasing the antibacterial responses and also helps in clearance of pathogen by increasing the TNF- α levels during *M. tb* infection. miR-146a and miR-155 are the most well-studied miRNAs in TB, which significantly influence the host-pathogen interactions [9].

miR-146a expression, driven by the transcription factor NF- κ B, also represses the mycobacteria-associated inflammatory response and helps in bacterial proliferation in RAW264.7 macrophages via IRAK-1/TRAF-6 pathway. miR-146a enhances mycobacterial survival in RAW264.7 macrophages via suppression of nitric oxide (NO) production [10]. miR-155 is induced upon infection of murine macrophages with mycobacteria and positively regulates the TLR signaling [5]. It is known to play a protective role against mycobacterial infections during the earlier stage of infection. On the other hand, it augments the survival of macrophages, thereby providing a niche for the mycobacterial replication. miR-155 levels in serum negatively regulate the activity of the NK cells. Increased levels of miR-155 lead to decreased

NO synthesis and enhance the mycobacterial load [11]. Targeting the miR-146a and miR-155 will be helpful in mycobacterial killing by increased NO production and increased NK cell activity. MyD88 signaling facilitates bacterial containment and is crucial for raising innate and adaptive immune response against *M. tb*. miR30a targets 3'-UTR of MyD88, thus inhibiting MyD88/TLR signaling in macrophages and promoting bacterial growth inside host cells. miR-27a targets 3'-UTR of TICAMI, a TLR3 adaptor, to inhibit TLR3 pathway [12]. miR-146a, an anti-inflammatory, regulates NF- κ B signaling and favors *M. tb* growth. miR-223 is significantly increased in macrophages infected with *M. tb* and negatively regulates NF- κ B signaling and pro-inflammatory cytokines [12].

7.2.2.3 Host miRNA Regulates Apoptosis in Tuberculosis

Apoptosis is an important mechanism to inhibit/kill intracellular mycobacteria. Host miRNAs that negatively regulate apoptosis are upregulated during *M. tb* infection. FOXO1, a positive regulator of apoptosis, is targeted by miR-582-5p in patients with active TB. Increased levels of miR-223 are associated with decreased apoptosis in *M. tb*-infected macrophages. Fas, a central component in apoptosis pathway, is targeted by let-7b-5p miRNA, thus inhibiting apoptosis in THP1 macrophages [12].

7.2.3 Host miRNAs as Diagnostic Biomarkers in Tuberculosis

An ideal biomarker for TB should have high sensitivity and specificity and the ability to be detected based on minimally invasive procedures [13]. A reliable biomarker for TB must be able to differentiate between healthy people and those with active TB and LTBI, as well as account for differences in expression because of variation in stages of infection and human-induced biases such as differences in age, sex, and comorbidities in patients [14]. Hence, exploring circulating host miRNA profiles has been considered a viable step in the direction of finding a potential biomarker for TB (Fig. 7.1).

7.2.3.1 Exosomal miRNAs as Biomarkers in Tuberculosis

Pulmonary TB accounts for around 80% of all TB forms, but tuberculosis meningitis (TBM) accounts for approximately 50% of TB-related mortality. There is a pressing need to have new and definitive biomarkers to improve TB diagnosis, especially for lethal forms such as TBM. Recently, Hu et al. showed that exosomal miRNAs could serve as reliable biomarkers for PTB and TBM combining microarray and electronic health record (EHR) data approaches [15]. They identified six differentially expressed human miRNAs in PTB and TBM patients compared to healthy controls. Integration of microarray and EHR data provided a better superior model in the differential diagnosis of PTB and TBM.

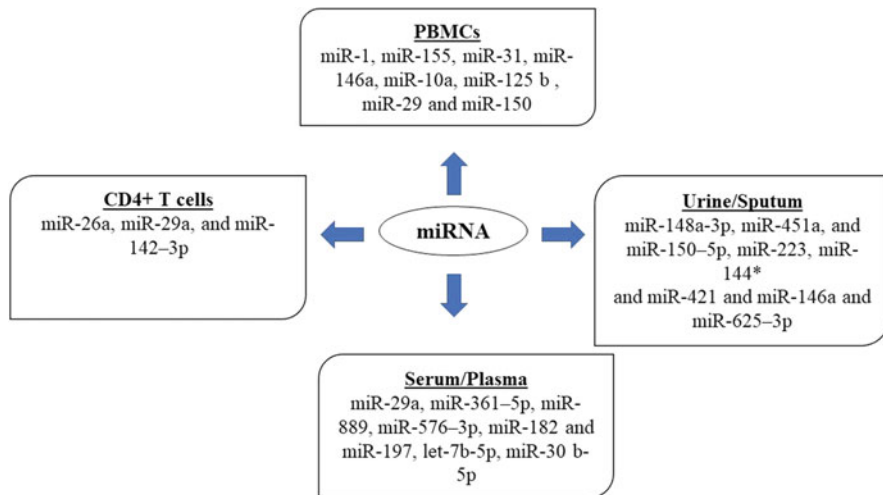


Fig. 7.1 A schematic illustration showing a list of miRNAs as biomarkers in tuberculosis for different biological specimen/samples

7.2.4 Role of miRNAs in Extrapulmonary TB

Extrapulmonary TB (EPTB) accounts for around 3% of TB cases. There is an urgent need to develop accurate and rapid diagnostic approaches to cure EPTB. Bone TB/skeletal TB is one form of EPTB, and spinal TB is one of the commonest bone TB found around the world. Yang et al. showed a negative correlation between host miRNA-155 and matrix metalloproteinase-13 (MMP-13) expression in patients with spinal TB-induced intervertebral disc, suggesting new approaches to treat and diagnose spinal TB [16].

TBM is one of the severe forms of TB of the central nervous system. Due to lack of accurate and efficient diagnostic methods, it is difficult to discriminate TBM from other similar diseases such as viral meningitis (VM). Interestingly, Pan et al. showed that four miRNAs (miR-126-3p, miR-130a-3p, miR-151a-3p, and miR-199a-5p) clearly show differential expression between TBM and VM patients, thus improving our understanding of pathogenesis between TBM and VM [17].

7.2.5 Host miRNA Profile in Response to TB Treatment

miRNA profiling post TB treatment helps in identifying drug-resistant mutants. In TB-susceptible patients (those who respond to anti-TB therapy), miR-320a is increased compared to drug-resistant TB patients. Notably, Wang et al. identified 37 upregulated and 63 downregulated miRNAs in humans cured of TB as compared to untreated TB patients [18]. Further, expression of miRNAs involved in Th1 response was modulated post TB treatment including miR-155 and miR-326.

miR-16 was expressed at lowest levels in MDR patients compared to all controls and healthy individuals.

7.2.6 Limitations with miRNA Profiling for Diagnostics in Tuberculosis

A plethora of studies have been conducted in order to explore the differential expression of miRNAs in different stages of TB infection. These studies show a diverse profile of upregulated and downregulated miRNAs between patients with LTBI and active TB infection. The inconsistencies in the results from the various studies are majorly due to the biases introduced as a result of methodological differences, the choice of sample, data analysis techniques, cut off criterion, etc. (i.e., human-induced biases) or because of the varying stages of TB infection in the samples taken [19]. Application of miRNAs as biomarkers in TB infection studies is often complicated by the ambiguity in the different stages of infection and the lack of differentiation between recent and long-standing latent TB infection and newly acquired or treated infections, which makes it difficult to compare the samples among the groups [19]. Moreover, the analysis of the same between the two sexes is done in an identical manner even though there is little to no similarity in the miRNA expression profiles of the two [20].

Previous studies have also shown various challenges associated with the delivery of miRNAs to their respective targets. These include a low half-life of miRNA in biological fluids; easy degradation and eradication of miRNA particles in vivo because of nuclease activity, phagocytosis, or renal clearance owing to their low molecular weight; inadequate penetration of the miRNA into tissues due to impediment caused by various physicochemical barriers; and intracellular disposition of miRNA for aggregating within endosomes of naked miRNA leading to inefficient gene-silencing or inducing immuno- or neurotoxicity [21, 22]. Recent studies have concluded that miRNAs are not as specific in their targeting action as they were previously assumed to be and can bind to complementary sequences as well as other targets having sequence similarity, which may lead to deleterious effects, ones that may even completely negate the therapeutic effects of miRNA [23].

Other potential drawbacks of miRNA studies in tuberculosis infection may arise depending on the techniques used in their profiling. NGS analysis or microarray techniques are most commonly used for searching possible miRNA biomarkers due to their advantages in allowing the identification and analysis of multiple targets [23]. However, both techniques have their own set of disadvantages as well, which may be extended to miRNA studies.

7.3 RNA-seq

At present, many technological platforms are available for sequencing, namely, Illumina, Roche 454, ABI SOLiD, Ion Torrent, and PacBio [24]. Illumina is the most commonly used platform because of its high sequence yield, but Ion Torrent and PacBio are gaining more popularity over Illumina platform due to their low cost and longer reads generation. Millions of reads are generated after RNA sequencing; hence, in-depth analysis is required. There is no single set of pipeline for the analysis as it depends on the availability of reference genome or transcriptome and also on the objective of the researcher [25]. Figure 7.2 summarizes an overview of the conventional RNA-seq data analysis workflow comprising both experimental design and computational analysis [25, 26]. Experimental workflow includes RNA extraction, reverse transcription polymerase chain reaction (RT-PCR), and cDNA library construction followed by sequencing and imaging. Computational analysis includes different software for each step varying from quality control to functional profiling. Programs are selected based on the aim of the work. Apart from command line tools, many other web-based, user-friendly pipelines such as Galaxy (galaxyproject.org), iDEP (ge-lab.org/idep), and UTAP (openpbs.org) are also available for RNA-seq dataset analysis, making it convenient for biologist not having bioinformatics expertise.

7.3.1 Differential Gene Expression Analysis

Differentially expressed gene (DEG) analysis is a process for identification of genes and comparison of their expression levels under different biological conditions, tissues, or stages of development.

7.3.2 Computational Analysis on RNA-seq

Recent studies have suggested that noncoding transcripts are not just noise or junks, they are biologically important; hence, they are now being implicated in human pathology [27]. The advancement of RNA-seq in recent years has facilitated the expression profiling of both long and short noncoding transcripts. Apart from expression profiling, quantification of exogenous RNA content, transcriptional elongation, and DNA variance can also be estimated using different RNA-seq approaches. ENCODE software tool helps in the sequencing data of long noncoding RNA through RNA-seq [28]. PathSeq tool is highly sensitive and specific and helps in discrimination of human from nonhuman sequences by use of data of transcriptome and whole-genome sequencing data [29]. Detection of virus can also be done through VirusSeq, ViralFusionSeq, and VirusFinder [28]. Software such as asSeq and AlleleSeq help in differentiation of expression of alleles by transcriptome and whole-genome sequencing data [28]. There are RNA-editing databases, such as DARNED, REDidb, dbRES, and RADAR, which are useful for removing

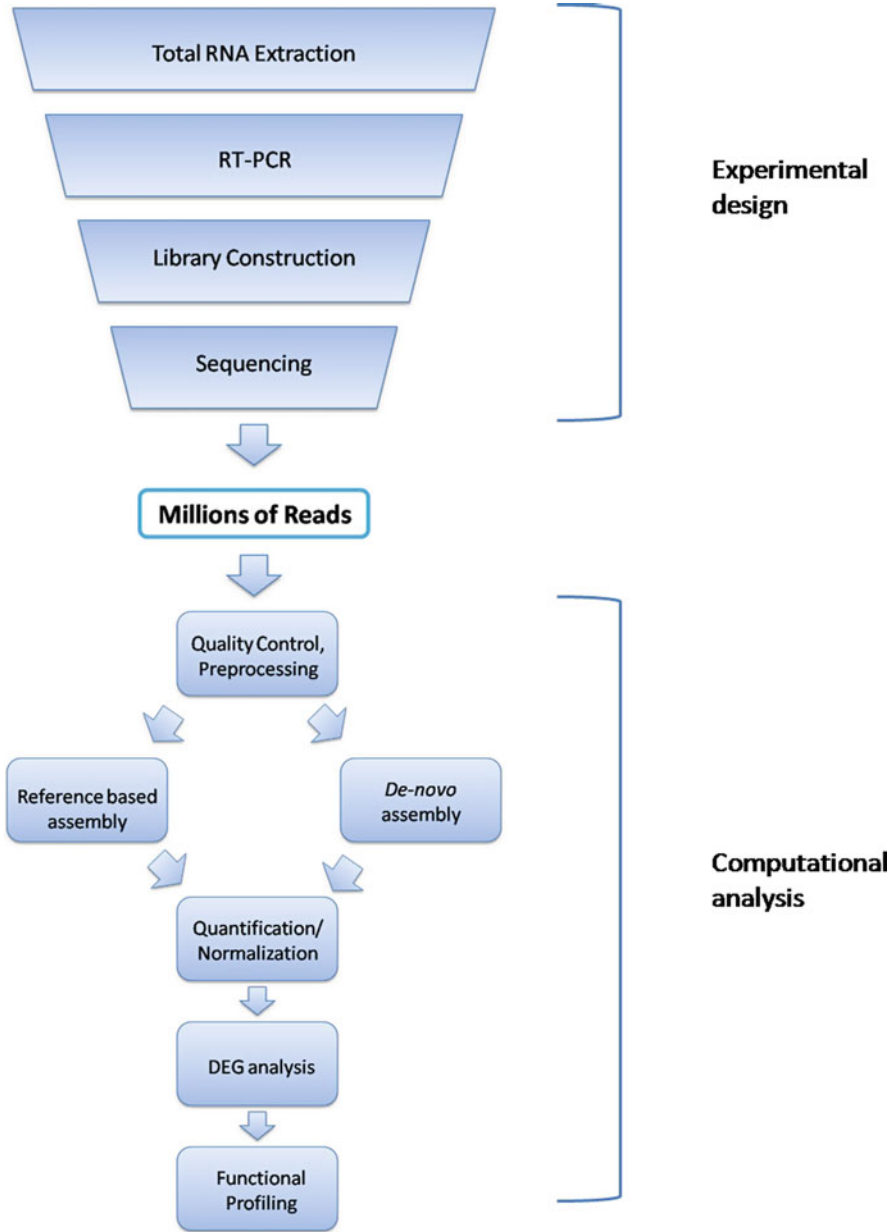


Fig. 7.2 Overview of RNA-seq data analysis workflow

RNA-editing sites [28]. Tools such as single-nucleotide variation quality, RNAmapper, RSMC, and SNPiR are useful for detection of DNA variations [28].

7.3.3 Advancements in RNA-seq

Although RNA-seq is a high-throughput technology, it has certain limitations. So some refinements have taken place in RNA-seq technology to overcome those limitations, some of which are as follows:

1. It is unable to determine polarity of RNA transcript, which is important for correct annotation of novel genes as it gives information related to the function of a gene, both at the level of RNA and the protein. To overcome this, single-strand RNA-seq technique was developed. Deoxy-UTP is incorporated in the second strand of cDNA. After this step, there is destruction of the uridine-containing strand which helps in the identification of the orientation of the transcript as shown in Fig. 7.3a [30].
2. RNA is extracted from the large population of cells in the sample. As a result, important differences between cells may be ignored, so a protocol of single-cell RNA-seq was developed as shown in Fig. 7.3b [31].
3. When host-pathogen interaction takes place, changes in the gene expression occur in both host and pathogen. So an advancement in RNA-seq has taken place which helps in studying the expression of the genes simultaneously in both host and pathogen known as dual RNA-seq, as shown in Fig. 7.3c [32].

7.4 Use of RNA-seq in Deciphering Complexities in *M. tb*

M. tb is able to disseminate successfully because of immune evasion depicted by it after infection in host, survival, and persistence. Successful pathogenesis and transmission of *M. tb* are mainly due to drug resistance, drug tolerance, and reactivation from dormancy. Recently developed technique of RNA-seq has provided in-depth information related to important attributes of *M. tb* such as dormancy, survival, infection, and immune evasion.

7.4.1 Dormancy

Fatty acid metabolism, rather than carbohydrate utilization, plays an important role in latency and dormancy. Fatty acid metabolism has been recently studied in depth through global transcriptome of *M. tb* by the use of strand-specific RNA-seq. *M. tb* bacilli were grown in media containing even-length long-chain fatty acids (LC-FAs) as the sole source of carbon and compared with those grown with dextrose supplementation [33]. The study revealed that there were a shift toward the glyoxylate cycle, increase in expression of several genes of reductive stress, and regulation like

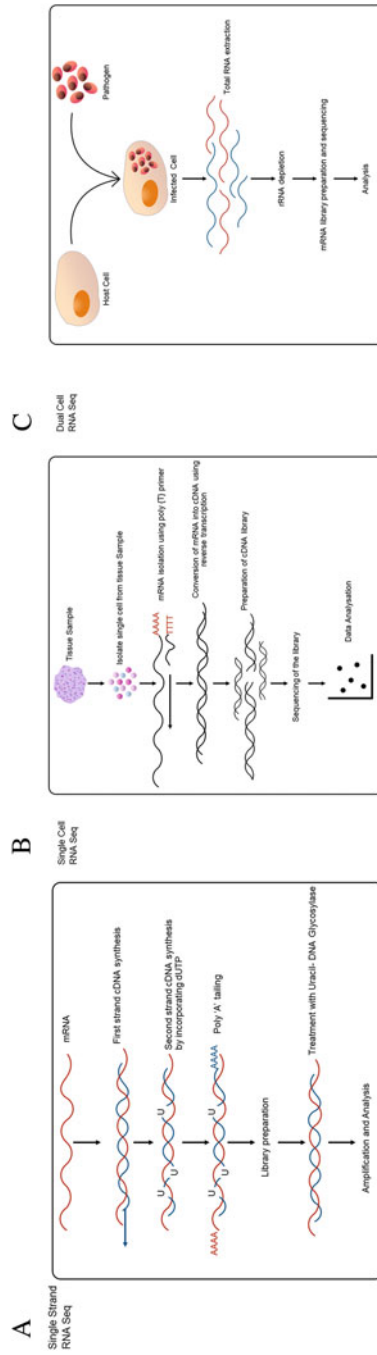


Fig. 7.3 Advancements in RNA seq. (a) Single-strand RNA seq. (b) Single-cell RNA seq. (c) Dual RNA-seq

whiB3, dosR, and Rv0081. There was also an upregulation of tRNA, which gave an attribute of drug tolerance by dormant bacterium. Lipid signature genes were also identified [33]. Another group did a transcriptome profiling using a mixture of cholesterol and palmitic, stearic, and oleic acids under aerobic and hypoxic conditions over three metabolic stages: exponential stage, stationary stage, and non-replicating persistence stage 1 of hypoxia [34]. They used cholesterol and long-chain fatty acids as a carbon source under aerobic and hypoxic conditions because sputum from TB patients showed the presence of lipid-rich environment where mycobacteria are residing and it mainly comprises cholesterol and palmitic, stearic, and oleic acids. The results revealed the differential expression of a core of 368 genes out of which 183 were downregulated and 185 were upregulated. This mainly helped in the induction of a machinery that leads not only to tolerance of drug but also in the maintenance of iron and production of sulfide, which is essential for enzymes and coenzymes necessary for redox balance and production of acetyl-CoA and methylmalonyl-CoA for de novo lipid biosynthesis [34]. Thus, these RNA-seq-based investigations have provided some clues related to target pathways in *M. tb* metabolism and hence paved the way to find alternate targets for therapy.

7.4.2 Survival

There are many hypothetical genes in *M. tb* which play an important role in survival of *M. tb* inside the host cell and a couple of genes from the host, playing roles in favor of the pathogen. As *M. tb* is an intracellular pathogen, it is able to face stress inside the macrophage and granuloma. It is thus necessary to study the response of *M. tb* to similar stresses which will help in identification of virulence factors and pathways necessary for the survival of *M. tb*. *M. tb* were exposed to stress such as oxidative, nitrosative, and alkylation stresses or mitomycin C-induced double-strand break in DNA. Nearly 700 DE genes were seen under nitrosative stress [35]. Mitomycin C stress mainly affects replication, repair, and recombination. Genes of the T7 secretion system (T7SS) and proline-glutamic acid/proline-proline-glutamic acid (PE/PPE) family responsible for the virulence of bacteria and survival were differentially expressed [35]. RNA-seq was also performed under various conditions of incubation, such as non-inhibitory to cidal condition based on the replication of mycobacteria or their killing profile [35]. There was an absence of replication of mycobacteria under inhibitory condition, and it lead to the expression of a unique transcriptome involving modulation of genes in response to stress, reprogramming in metabolic pathways, oxidative stress, response to dormancy, respiration, and virulence [35]. RNA-seq also gave transcription pattern which involved remodeling of cell wall, leading to increase in infection, resistance to antibiotics, and immune evasion.

7.4.3 Infection

Studies related to host-pathogen interactions in TB are important for the development of new strategies to control *M. tb*. In one of these studies, transcriptome analyses were carried out in macrophages which were infected with H37Rv (Rv), a virulent strain of *M. tb*, or H37Ra (Ra), an attenuated strain. Seven hundred fifty differentially expressed genes (DEGs) were identified and analyzed, of which *solute carrier family 7 member 2 (SLC7A2)* was more suppressed in Rv-infected macrophages as compared to Ra-infected macrophages [36]. SLC7A2 transporter is required by M1 macrophages to produce NO. M1 macrophages help in removal of bacteria because they produce pro-inflammatory cytokines and NO. Thus, intracellular survival of *M. tb* is also regulated by induction of SLC7A2.

Different cell lineages of the host showed differentially expressed genes as well as change in the expression of *M. tb* genes in pathogenesis through dual RNA-seq. Three thousand four hundred fifty-three genes were upregulated and 3119 downregulated in alveolar macrophages (AMs) compared with uninfected macrophages. The NRF2 pathways were upregulated in AMs infected with *M. tb*, while expression of pro-inflammatory cytokines and molecules for prevention of inflammation-associated intracellular damage was lower than infected interstitial macrophages (IMs). These results revealed that the infected AMs were more permissive for bacterial growth. In IMs, 3614 genes were upregulated, and 3298 genes were downregulated [32]. Transcriptional profiling of uninfected and infected IM populations depicted that pro-inflammatory pathways such as activation of NF- κ B and immune response through Th1 pathway got induced. Genes responsible for adhesion and chemotaxis were upregulated in IM-infected macrophages which recruited the other host cells to granuloma from peripheral blood. There was an induction of type II toxin-antitoxin-chaperon (higB1-higA1-secB) under the conditions of stress and hence played role in bacterial persistence [32].

Acquisition of nutrient in the presence of environment of the host is critical for the intracellular pathogens to survive. For the quantification of interactions in metabolic pathways between the host and *M. tb*, dual RNA-seq was performed. This resulted in identification of alterations in metabolic pathways which were specific to infection. The in silico data suggested that *M. tb* was able to consume 33 different nutrients during initial phase of macrophage infection, which is utilized by the tubercle bacilli for the generation of energy for its intracellular growth [37].

PathSeq tool has helped in the discovery of a novel transcriptional machinery for remodeling of cell wall of mycobacteria when *M. tb* infects alveolar macrophage. It has been shown that MadR modulates two mycolic acid desaturases *desA1/A2* for promotion of remodeling of cell wall when *M. tb* infects macrophages and helps in mycolate biosynthesis upon entering dormancy [38]. Thus, RNA-seq has provided an insight that disrupting MadR is lethal and can be an antitubercular target for initial and later stages of infection.

7.5 Application of RNA-seq in Developing Better Diagnosis and Treatment of Tuberculosis

It is important to understand response against tuberculosis (TB) infection for diagnosis of people having an active or a latent TB infection (LTBI). It is also important to identify LTBI individuals who are at greater risk of active TB development. The RNA-seq analysis was done on samples from Galicia (Spain) and Mozambique where more TB is prevalent [39]. The results show high expression of genes that code for immunoglobulin chains. There was an increase in the expression of immunoglobulin receptors such as FCGR1A and FCGR1B. Genes responsible for T cell regulation such as programmed cell death 1 ligand 1 (CD274) and 2 (PDCD1LG2) were also upregulated in samples obtained from active TB patients. Authors have identified two profiles within LTBI contacts. Transcriptional profiles of the 77.8% subgroup were similar to that of uninfected contacts, whereas the second subgroup (22.2%) showed transcription profiles similar to confirmed TB patients. Thus, these kinds of studies can be useful to study LTBI contacts at risk, progressing to active TB without any clinical follow-ups. Single-cell RNA-seq datasets obtained from peripheral blood monocytes of healthy control, LTBI and active tuberculosis revealed natural killer (NK) cell subset (CD3⁻CD7⁺GZMB⁺) got depleted in patients suffering from TB. Hence, it can serve as a biomarker for diagnosis of TB which will help in distinguishing LTBI from healthy control [40].

Antibiotics have a property to either stimulate or repress gene expression in bacteria. Through transcriptional profiling mode of action of various drugs, their efficacy and effect on the metabolism of *M. tb* can be deduced. Through this technique, new insights for known antibiotics that are currently useful for the treatment of tuberculosis can be obtained. This will also help in identification of mode of action and prediction of various targets of new drugs. This technique also provides information related to changes in transcriptome profile after treatment with antibiotic [41]. It also provides information related to capability of bacteria to escape the effect of antibiotics. So it will be helpful in identification of genes responsible in adaptive responses and tolerance to drugs. Transcriptome profiling helps in identification of genes differentially expressed between drug-sensitive *M. tb* and multidrug-resistant clinical isolates. Clinical strains of *M. tb* are different from H37Rv, more specifically clinical drug-resistant strains. By the help of transcriptome sequencing multidrug-resistant strains of *M. tb*, attributes of resistance and virulence were depicted [42]. There was an enrichment in biosynthesis of arginine, biosynthesis of fatty acid, and pathways of metabolism which was depicted through the DEGs of MDR strains. Type VII secretion system was downregulated in MDR strains.

7.6 Conclusions

Characteristic miRNAs, expressed at different stages of tuberculosis, might give useful insights on the nature of immune responses generated against this elusive pathogen. The advantage of using miRNA profile as a diagnostic biomarker is their

stability in the circulating body fluids. However, inconsistent results due to variability in miRNA expression levels, lack of optimal experimental models, and correlation of the results with the actual pathologies limit their use, and further validation is required to conclude clearly on the role of miRNAs as biomarkers in TB diagnosis and treatment. Nevertheless, we can conclude that (1) miRNAs were differentially expressed between active TB and LTBI patients, (2) the results of in vitro and ex vivo studies can vary considerably and can contradict each other, (3) PTB and EPTB can provide different miRNA signatures, and (4) treatment to TB patients highlights changes in miRNA profiles, indicating their potential use as a measure of efficacy to anti-TB therapy.

The advent of next-generation sequencing technologies that enabled transcriptomic investigations revolutionized the field of mycobacterial research at an unprecedented scale. However, the application of the transcriptomics-based investigation methods in mycobacterial research is still in its infancy and is yet to reach its zenith in the years to come. Once the full potential of these methods is realized and applied, the field will definitely be endowed with the information and insights on enigmatic mycobacterial lifestyle in its favored niche that is still largely elusive.

References

1. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57–63
2. WHO. Global Tuberculosis Report (2020) Available online at: https://www.who.int/tb/publications/global_report/en/. Accessed 14 Oct 2020
3. Chakrabarty S et al (2019) Host and MTB genome encoded miRNA markers for diagnosis of tuberculosis. *Tuberculosis* 116:37–43
4. Gan J, Qu Y, Li J, Zhao F, Mu D (2015) An evaluation of the links between microRNA, autophagy, and epilepsy. *Rev Neurosci* 26(2)
5. Wang J, Yang K, Zhou L, Minhaowu, Wu Y, Zhu M et al (2013) MicroRNA-155 promotes autophagy to eliminate intracellular mycobacteria by targeting Rheb. *PLoS Pathog* 9(10): e1003697
6. Kim JK, Yuk JM, Kim SY, Kim TS, Jin HS, Yang CS et al (2015) MicroRNA-125a inhibits autophagy activation and antimicrobial responses during mycobacterial infection. *J Immunol* 194(11):5355–5365
7. Naqvi AR, Fordham JB, Nares S (2015) miR-24, miR-30b, and miR-142-3p regulate phagocytosis in myeloid inflammatory cells. *J Immunol* 194(4):1916–1927
8. Li H, Jiang T, Li MQ, Zheng XL, Zhao GJ (2018) Transcriptional regulation of macrophages polarization by microRNAs. *Front Immunol* 9:1175
9. Sabir N, Hussain T, Shah SZA, Peramo A, Zhao D, Zhou X (2018) miRNAs in tuberculosis: new avenues for diagnosis and host-directed therapy. *Front Microbiol* 9: 602
10. Li M, Wang J, Fang Y et al (2016) microRNA-146a promotes mycobacterial survival in macrophages through suppressing nitric oxide production [published correction appears]. *Sci Rep* 6:24555
11. Qin Y, Wang Q, Zhou Y, Duan Y, Gao Q (2016) Inhibition of IFN- γ -induced nitric oxide dependent antimycobacterial activity by miR-155 and C/EBP β . *Int J Mol Sci* 17(4):535
12. Yang T, Ge B (2018) miRNAs in immune responses to *Mycobacterium tuberculosis* infection. *Cancer Lett* 431:22–30

13. Correia CN, Nalpas NC, McLoughlin KE, Browne JA, Gordon SV, MacHugh DE, Shaughnessy RG (2017) Circulating microRNAs as potential biomarkers of infectious disease. *Front Immunol* 8:118
14. Pedersen JL, Bokil NJ, Saunders BM (2019) Developing new TB biomarkers, are miRNA the answer? *Tuberculosis (Edinb)* 101860:118
15. Hu X, Liao S, Bai H, Wu L, Wang M, Wu Q, Zhou J, Jiao L, Chen X, Zhou Y, Lu X, Ying B, Zhang Z, Li W (2019) Integrating exosomal microRNAs and electronic health data improved tuberculosis diagnosis. *EBioMedicine* 40:564–573
16. Yang C, Shi Z, Hu J, Wei R, Yue G, Zhou D (2019) miRNA-155 expression and role in pathogenesis in spinal tuberculosis-induced intervertebral disc destruction. *Exp Ther Med* 17 (4):3239–3246
17. Pan L, Liu F, Zhang J, Li J, Jia H, Huang M, Liu X, Chen W, Ding Z, Wang Y, Du B, Wei R, Sun Q, Xing A, Zhang Z (2019) Genome-wide miRNA analysis identifies potential biomarkers in distinguishing tuberculous and viral meningitis. *Front Cell Infect Microbiol* 9:323
18. Wang C, Yang S, Liu CM, Jiang TT, Chen ZL, Tu HH, Mao LG, Li ZJ, Li JC (2018) Screening and identification of four serum miRNAs as novel potential biomarkers for cured pulmonary tuberculosis. *Tuberculosis (Edinb)* 108:26–34
19. Ruiz-Tagle C, Naves R, Balcells ME (2020) Unraveling the role of microRNAs in Mycobacterium tuberculosis infection and disease: advances and pitfalls. *Infect Immun* 88(3):e00649–e00619
20. Corral-Fernández NE, Cortes-García JD, Bruno RS, Romano-Moreno S, Medellín-Garibay SE, Magaña-Aquino M, Salazar-González RA, González-Amaro R, Portales-Pérez DP (2017) Analysis of transcription factors, microRNAs and cytokines involved in T lymphocyte differentiation in patients with tuberculosis after directly observed treatment short-course. *Tuberculosis (Edinb)* 105:1–8
21. Judge AD, Sood V, Shaw JR, Fang D, McClintock K, MacLachlan I (2005) Sequence-dependent stimulation of the mammalian innate immune response by synthetic siRNA. *Nat Biotechnol* 23(4):457–462
22. Dua K, Hansbro NG, Foster PS, Hansbro PM (2017) MicroRNAs as therapeutics for future drug delivery systems in treatment of lung diseases. *Drug Deliv Transl Res* 7(1):168–178
23. Mehta M, Chellappan DK, Wich PR, Hansbro NG, Hansbro PM, Dua K (2020) miRNA nanotherapeutics: potential and challenges in respiratory disorders. *Future Med Chem* 12 (11):987–990
24. Oszolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 12:87–98
25. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol* 17:13
26. Han Y, Gao S, Muegge K, Zhang W, Zhou B (2015) Advanced applications of RNA sequencing and challenges. *Bioinform Biol Insights* 9:29–46
27. Jathar S, Kumar V, Srivastava J, Tripathi V (2017) Technological developments in lncRNA biology. *Adv Exp Med Biol* 1008:283–323
28. Han L, Vickers KC, Samuels DC, Guo Y (2015) Alternative applications for distinct RNA sequencing strategies. *Brief Bioinform* 16:629–639
29. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak R, Getz G, Meyerson M (2011) PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* 29(5):393–396
30. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitch S, Hans Lehrach H, Soldatov A (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* 37(18):e123
31. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6:377–382

32. Pisu D, Huang L, Grenier JK, Russell DG (2020) Dual RNA-Seq of Mtb-infected macrophages in vivo reveals ontologically distinct host-pathogen interactions. *Cell Rep* 30:335–350.e334
33. Rodriguez JG, Hernandez AC, Helguera-Repetto C, Aguilar Ayala D, Guadarrama-Medina R, Anzola JM, Bustos JR, Zambrano MM, Gonzalez YMJ, Garcia MJ, Del Portillo P (2014) Global adaptation to a lipid environment triggers the dormancy-related phenotype of *Mycobacterium tuberculosis*. *mBio* 5:e01125–e01114
34. Aguilar-Ayala DA, Tillemann L, Van Nieuwerburgh F, Deforce D, Palomino JC, Vandamme P, Gonzalez YMJA, Martin A (2017) The transcriptome of *Mycobacterium tuberculosis* in a lipid-rich dormancy model through RNAseq analysis. *Sci Rep* 7:17665
35. Namouchi A, Gomez-Munoz M, Frye SA, Moen LV, Rognes T, Tonjum T, Balasingham SV (2016) The *Mycobacterium tuberculosis* transcriptional landscape under genotoxic stress. *BMC Genomics* 17:791
36. Lee J, Lee SG, Kim KK, Lim YJ, Choi JA, Cho SN, Park C, Song CH (2016) Characterisation of genes differentially expressed in macrophages by virulent and attenuated *Mycobacterium tuberculosis* through RNA-Seq analysis. *Sci Rep* 9:4027
37. Zimmermann M, Kogadeeva M, Gengenbacher M, McEwen G, Mollenkopf HJ, Zamboni N, Kaufmann SHE, Sauer U (2017) Integration of metabolomics and transcriptomics reveals a complex diet of *Mycobacterium tuberculosis* during early macrophage infection. *mSystems* 2
38. Peterson EJ, Bailo R, Rothchild AC, Arrieta-Ortiz ML, Kaur A, Pan M, Mai D, Abidi AA, Cooper C, Aderem A, Bhatt A, Baliga NS (2009) Path-seq identifies an essential mycolate remodeling program for mycobacterial host adaptation. *Mol Syst Biol* 15:e8584
39. Estevez O, Anibarro L, Garet E, Pallares A, Barcia L, Calvino L, Maueia C, Mussa T, Fdez-Riverola F, Glez-Pena D, Reboiro-Jato M, Lopez-Fernandez H, Fonseca NA, Reljic R, Gonzalez-Fernandez A (2020) An RNA-seq based machine learning approach identifies latent tuberculosis patients with an active tuberculosis profile. *Front Immunol* 11:1470
40. Cai Y, Dai Y, Wang Y, Yang Q, Guo J, Wei C, Chen W, Huang H, Zhu J, Zhang C, Zheng W, Wen Z, Liu H, Zhang M, Xing S, Jin Q, Feng CG, Chen X (2020) Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis. *EBioMedicine* 53:102686
41. Briffotiaux J, Liu S, Gicquel B (2019) Genome-wide transcriptional responses of mycobacterium to antibiotics. *Front Microbiol* 10:249
42. Tang J, Liu Z, Shi Y, Zhan L, Qin C (2020) Whole Genome and transcriptome sequencing of two multi-drug resistant *Mycobacterium tuberculosis* strains to facilitate illustrating their virulence in vivo. *Front Cell Infect Microbiol* 10:219



Microarrays: A Road Map to Uncover Host Pathogen Interactions

8

Heerak Chugh, Gagan Dhawan, Ramesh Chandra, and Uma Dhawan

Abstract

Microarray is among the recent development of molecular biology that has been applied to various fields of biological research. It is a high-throughput technology perfected over the years. Briefly, microarray can be defined as a lab-on-a-chip technique that collectively contains spots of thousands of biological material like nucleic acid or protein on a silicon or glass chip. These spots with which microarray are produced can be customized according to the experiment's requirements and primarily depend on the available sequence of the biological material of the organism. The basic principle that the microarray is based on is similar to the hybridization technique; therefore, a test sample containing the appropriate biomaterial is prepared and is hybridized with the microarray. The hybridization with complement biomaterial generates signals that can be read and analyzed to assess the expression of genes or proteins in the test samples. In this

H. Chugh

Department of Biomedical Science, Acharya Narendra Dev College, University of Delhi, Kalkaji, New Delhi, India

Department of Chemistry, University of Delhi, Delhi, India

G. Dhawan

Department of Biomedical Science, Acharya Narendra Dev College, University of Delhi, Kalkaji, New Delhi, India

R. Chandra

Department of Chemistry, University of Delhi, Delhi, India

U. Dhawan (✉)

Department of Biomedical Science, Bhaskaracharya College of Applied Sciences, University of Delhi, Dwarka, New Delhi, India

e-mail: uma.dhawan@bcas.du.ac.in

chapter, we discuss the principle, origin, and application of microarrays to decipher host-pathogen interactions.

Keywords

Microarrays · Gene expression · Genetic profiling · Genetic diagnosis · Gene chip

8.1 Introduction

Among the many high-throughput tools invented, microarrays have emerged as a powerful tool for analyzing the whole genome, transcriptome, and proteome of an organism. On a general note, the technique uses two-dimensional substrate made of either silicon or glass, carrying probes of biological materials such as DNA, RNA, or protein at a known coordinate on the substrate. A target sample, when applied to the array, hybridizes with its complementary counterpart to give off a signal which, when analyzed, reveals the identity and expression level of the biological material. In the current times, microarrays can use various biological materials like DNA, mRNA, microRNA, and proteins; however, the original microarray technology was invented to evaluate the expression of genes quantitatively. The microarray technology was made possible because of the success of large-scale genome-sequencing projects of various organisms like yeast, mouse, rat, and humans. According to the central dogma, DNA molecules that make up a genetic sequence are transcribed into mRNA molecules which are then translated into proteins. In the primary and original microarrays, DNA molecules from a source are fixed to a solid substrate. Total mRNA from a sample whose genetic expression has to be analyzed is reverse-transcribed into cDNA, digested, and labeled with Cy3 (green) and Cy5 (5) fluorescent dyes. These probes are then allowed to interact with the DNA fragments on the microarray; hybridization among complementary sequence gives a fluorescent emission upon excitation which is then analyzed using a software to provide the identity and the level of expression of the gene at the particular spot [1] (Fig. 8.1). Instruments with fluorescence scanners such as photomultiplier tubes or charged coupled devices (CCD) are used to detect the microarray post-hybridization to produce two-dimensional images [2]. The image produced represents the surface of the microarray substrate from 0 to 65,536 intensity values for each spot that corresponds to its expression level. The numerical data affiliated with the expression of each gene is analyzed and modeled using scatter plots, self-organizing maps, and other software tools that help deduce relevant information [3]. A microarray permits the user to explore the high-throughput and parallel differential global gene expression; the technique takes precedence over other methodologies like northern blotting and polymerase chain reaction which are limited to evaluating the expression of a handful number of genes at a time [4]. However, the production of microarrays is based on prerequisite information about the genome sequence and information restricting its use and limiting its effectiveness [5]. Years of modifications have equipped microarray technology to be applied to processes other than global gene

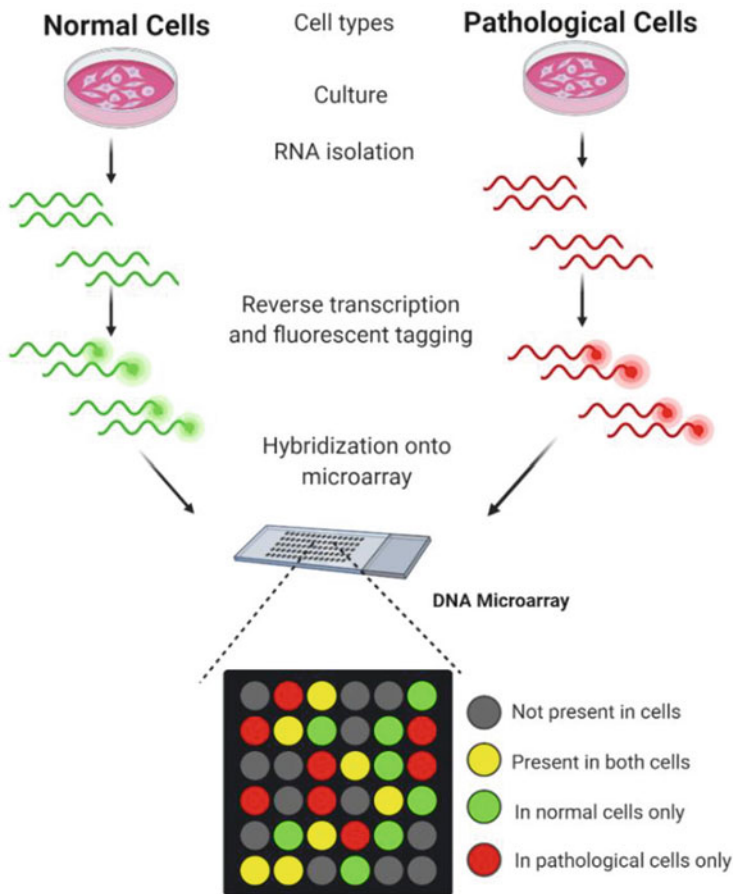


Fig. 8.1 DNA microarray: The figure depicts the basic methodology to formulate an oligonucleotide microarray to differentiate the changes in gene expression of a pathological cell from a normal cell (created with Biorender.com)

expression profiling. These modifications and the various applications that the microarrays have been successfully employed are discussed in the following sections of this chapter.

8.2 Origin of Different Types of Microarray Technologies

The history of the modern-day microarray technology dates back to mid-1970s when Grunstein and Hogness developed a new method called colony hybridization [6]. In this method, *Escherichia coli* was transformed with random DNA sequences and plated on the nitrocellulose filter membrane. Post colony formation, the bacterial cells were lysed, and the DNA that each colony carried was denatured and fixed,

giving rise to a collection of DNA sequences or DNA spots which were further hybridized with and screened for radiolabeled complementary RNA sequences. The concept led to the development of mechanical devices that could print colonies in microtiter plates which were transferred to Whatman filter paper in an orderly fashion to produce reusable arrays. The arrays developed using this technique were applied to clone a gene of interest, map the physical distance between genes, and analyze differential expression. However, the vast applicability was handicapped by the slow speed of production and operation and induced human errors. Thus, the production of filter arrays was automated to increase speed and accuracy by Hans Lehrach [7]. In 1995, Pan Brown and Ron Davis at Stanford University used the robotics system to spot cDNA molecules from *Arabidopsis thaliana* on glass slides, developing the concept of “gene chips” which were further hybridized with fluorescently labeled sample probes [8]. In 1996, Derisi used polylysine-coated glass microscopic slides for a better DNA binding to analyze the patterns in gene expression in human melanoma cell lines [9]. On a different approach, in 1994, Steve Fodor developed DNA microarrays by directly synthesizing DNA oligonucleotides on a solid substrate under a brand that now has become associated explicitly with DNA microarrays—Affymetrix [10]. Affymetrix technology was no doubt better than the labor-intensive spotted array technology as it requires lesser number of starting materials to construct arrays; however, the production cost for custom arrays with smaller volumes was high because of the use of distinctive photolithographic masks. Flexibility to the technique was added by adopting the use of a digital light processor to produce virtual masks, giving rise to a maskless approach [11]. In another approach to reduce the costs of microarray fabrication, inkjet printing was employed to dispense biomolecules on a solid substrate. The process was first exploited by Blanchard et al. in 1996 wherein the chemical reagents to synthesize DNA probes were dispensed in a hydrophilic region of pre-patterned silicon dioxide substrate to produce high-density custom arrays in low volume [12]. In another intriguing advance, arrays using fiber optics to detect fluorescently labeled microspheres or beads carrying DNA molecules were developed by Illumina [13]. A mixture of these beads carrying different DNA molecules was randomly applied to a fiber-optic substrate, followed by the decoding of the fluorescent labeling on each bead. The technique was further modified to detect fluorescently labeled DNA molecules on the beads instead and use of etched glass surface instead of fiber-optic arrays [7]. These microspheres were also employed in a three-dimensional suspension array instead of being fixed in a two-dimensional microarray. These beads are labeled with specific fluorochromes and probes; the suspension mixture of randomly assembled beads is then assorted using flow cytometry [14]. Electric fields have also been used efficiently for the fabrication of electronic microarrays; they are electrode-based arrays that can be influenced by modifying the strength and nature of electric fields [15]. The positive electronic fields are used to place the negatively charged nucleic acid at an electronically activated spot on the array until all the spots are probed with different nucleic acid molecules. Once the array is complete and ready to use, sample target nucleic acid probes are applied, and the hybridization that may

occur is scanned and analyzed employing fluorescent reporters. From the brief history narrated above, it is evident that the production and implementation of microarrays has advanced and evolved into more economical, concise, and efficient technology over the years for their successful application in various fields.

8.3 Applications of Microarrays

With the advance of microarray technologies, proteins, cell lysate, and tissues samples have also been immobilized on the array substrate, increasing the number of applications that can be performed using this technology. On a brief note, microarray technology is used for gene expression analysis/profiling; point mutation analysis; gene expression analysis in cancer, metabolism, drug discovery and development, toxicity, neurological disorders, and infectious diseases; diagnostics of various infections and diseases; and evaluation of the changes in gene expression in response to drugs and pathogen invasion [16].

8.3.1 Genetic Profiling

Although sequencing by hybridization has now been replaced by other methods, it was among the earliest applications for which microarrays were employed. It is possible to decipher the sequence of a target DNA molecule by analyzing its pattern of hybridization with an array consisting of probes of the same length with different nucleotide combinations as the target DNA molecule [17]. The concept of genome expression profiling that followed the production of microarrays carrying genomes of diverse organisms and diverse ethnicity among humans led to the unwinding of exciting facts that shaped the attributes we know about the effect of genetic diversity on phenotype. The changes in genome content of an organism through nucleotide insertions and deletions over time can be ascertained through the use of microarray. In a study, it was deciphered that 37 insertions and 30 deletions took place to make up the genomic content of *E. coli* MG1655 [18]. Microarrays have also explained the genetic basis of differences among two different species; scientists have compared and profiled genes like SMAD1, GTF21, and TWIST1 among many others to be differentially regulated between the cortical tissues of human and chimpanzee brain, providing insights into the evolution of the human brain [19]. Thus, microarrays can be used both within a species and between two different species to chart out the genetic changes that led to the modern-day evolved phenotype. Similarly, genetic mutation, polymorphism, and allelic variation in a population can also be analyzed using DNA microarrays. This approach has also made gene profiling among various cells and tissues of human body possible to trace the differential expression of genes and intergenic regions in different cells and tissue [20]. Considerable use of microarrays has been seen in detecting differential gene expression associated with various diseased states and infections as any disease phenotype is directly linked to chromosome/gene abnormality and/or modifications in gene expression. In addition

to DNA microarrays, cell and tissue microarrays, protein arrays or chips, and some other specialized arrays are also being developed and employed.

8.3.2 Drug Discovery and Development

In a more preliminary approach to drug discovery and development, enzymes and/or proteins involved in the pathological process of a disease or a therapeutic response are screened for drugs that have previously shown interaction with similar targets. However, the approach is limited by the number of targets available for a specific pathway involved in the pathophysiology of disease; DNA microarrays and protein chips can identify and validate novel targets for therapeutic interventions. Targeting the expression of genes that either cause the illness or manifest the symptoms is another approach for therapeutic intervention. The aim is to select a few candidate genes which maybe increased or decreased disease-specific and highly specific for cell or tissue. These preliminary gene candidates are then further scrutinized in secondary and tertiary screenings to be validated as potential therapeutic targets [21]. Genes for p-glycoprotein-like protein, ABC1 transporter, guanine nucleotide-binding protein, and a few other genes have been identified as potential drug targets for visceral leishmaniasis [22]. Microarrays can also be utilized to elucidate the therapeutic, resistance, and toxic mechanisms in response to drug treatment. For instance, a microarray analysis of VCR-resistant MG63 osteosarcoma cells when compared with corresponding nonresistance cell lines reported differential expression of 1300 genes which may be involved in conferring resistance to MG63 cell line [23]. Similarly, comparison among the differential expression of normal, diseased, and drug-treated cells may provide insights into the drug's mechanism [24]. The use of microarrays in drug discovery and development has led to the production of some specific microarrays as well, for instance, the development of a G protein-coupled receptors microarray, since they are a fundamental class of therapeutic targets [25].

8.3.3 Application in Microbiology

Microarrays have found a significant application in the field of microbiology because of their low-cost, accurate, and reliable results. They have significantly improved our knowledge about the pathogenic gene expression, underlying mechanisms, effect of environmental stimuli, identification of new pathogens and new strains, antimicrobial therapeutics, antimicrobial resistance, and host-pathogen interactions [26].

DNA microarrays provide a more sensitive, rapid, and efficient way of pathogen detection in comparison to more conventional approaches like culture-based methods, ELISA, and PCR [27]. Scientists have developed common microarray panels to identify different pathogens, for instance, multi-pathogen identification (MPID) microarray for the identification of 18 pathogenic strains of prokaryotes, eukaryotes, and viruses [28], for the rapid identification of causative agent of purulent meningitis [29], viral meningitis, and encephalitis [30] and many more.

Microarrays have also been successfully employed for microbial typing to define pathogens at subspecies level by characterizing the genetic changes (insertions, deletions, substitutions, or rearrangements) for several pathogens, including bacteria like *Salmonella enterica*, *E. coli*, and *Yersinia pestis* [31], and influenza virus [32]. Similarly, microarrays have enabled researchers to identify and characterize rapidly evolving superbugs like MRSA to decipher the evolution in their virulence and resistance factors at a global level [33, 34]. Further, specific microarrays have been developed for genotyping the antimicrobial drug resistance in several pathogenic strains. Many microarrays have been produced carrying probes with antibiotic resistance genes against certain antibiotics to test specific strains; for instance, NCBI was successful in producing microarray with oligonucleotide probes of 775 antibiotic resistance genes that confer resistance against some of the commonly used antibiotics, such as chloramphenicol, glycopeptides, aminoglycosides, and sulfonamides [35]. The use of such microarrays has significantly improved the screening of resistant strains in clinical and environmental settings along with an enhanced understanding of the basis of resistance development. DNA microarrays have also evaluated the change in the genetic expression of pathogens in response to therapeutic interventions to provide significant insights into the mechanism of action of the drug as well as other changes in their environment [36–38].

8.3.4 Host-Pathogen Interactions

When a pathogen invades a host organism, the interaction between the two brings about a plethora of changes in both the pathogen and the host; to identify these changes, DNA and protein microarrays have been used (Fig. 8.2). The development and production of microarrays of several pathogenic microorganisms like *Mycobacterium tuberculosis* [39], *Neisseria meningitidis* [40], *Plasmodium* sp. [41], HIV [42], and SARS-CoV-2 [43, 44] among others and mammalian microarrays have made it possible to study host-pathogen interaction at transcriptome and proteome levels. In the remainder of this chapter, we are going to explore the concepts of host-pathogen interaction with suitable examples involving pathogenic virus. In the most basic idea, the change in the host gene expression is evaluated after infection with the pathogenic virus [45]. The genetic expression in the host can be assessed at different stages of infection, in different age groups, in individuals from various regions, and other varying host factors to identify and characterize genes involved in disease pathogenicity and therapeutics. The analysis of gene expression as a response to variola infection (smallpox virus) in PBMCs of 22 cynomolgus macaques using human DNA microarrays with 18,000 unique genes revealed the upregulation of IFN genes, cell proliferation genes, unmodulated TNF- α /NF κ B pathway, and decreased lymphocyte gene expression [46]. The study has provided a temporal correlation between the genetic expressions in the blood monocytes and the clinical signs showed by the animals, offering new insights into development, pathogenicity, and potential therapeutic targets during different stages of infection.

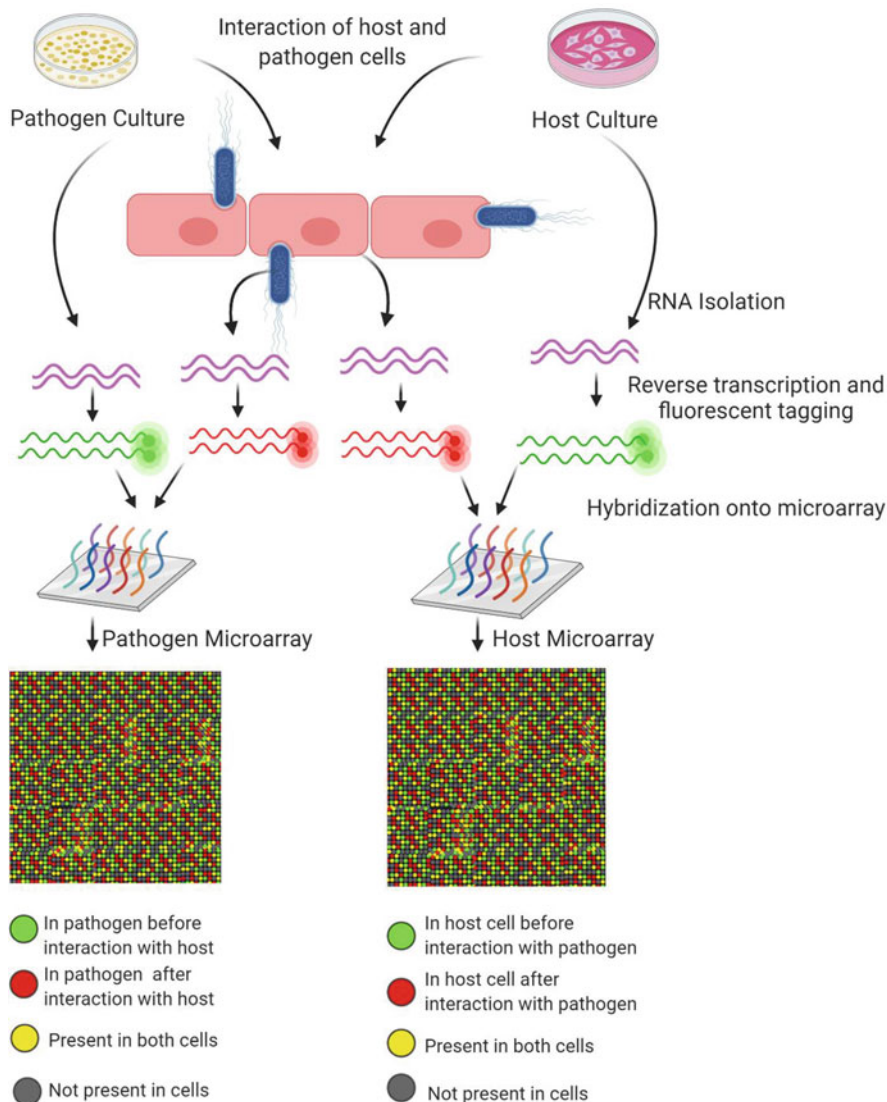


Fig. 8.2 Host-pathogen interaction: the figure depicts a sequential methodology for studying the changes in gene expression in host and pathogen cell after their interaction using microarrays (created with Biorender.com)

Similarly, genetic microarrays over the years have generated large sets of data that have unified the gene signatures linked to various pathways involved in disease progression upon HIV infection [47]. CD4+ T cells, which are the primary site for HIV infection are consistently depleted in numbers in lymphoid tissues and circulation with an increase in expression of genes that are related to cell cycle regulation,

interferon, and apoptosis [48]. The use of proteome microarrays has also been successful in studying the host immune response upon viral infection. In this concept, the sera from an infected individual are compared to noninfected individual (or control). They are applied to a proteome microarray consisting of potential viral targets that may be immunogenic and elicit an immune response in the host body. For instance, a humoral response (IgG and IgM profiles) was characterized for COVID-19 patients in comparison to influenza and non-influenza candidates using a SARS-CoV-2 microarray. It was discovered that 27 IgM and 9 IgG antigens were specific to the immune response in COVID-19 patients including spike protein eliciting both IgM and IgG response while nucleocapsid and protease-ORF1ab mounting mainly IgG response [49]. In convalescent COVID-19 patients, the immune response was stronger for nucleocapsid, spike, and ORF9b proteins; however, a strong IgG binding for nucleocapsid was observed in control samples as well, owing to conserved motifs throughout the coronavirus family [50]. In the same study, it was also deduced that the host response to different versions of spike protein was uncorrelated. Subsequently, a microarray was developed for epitope mapping of the spike protein of SARS-CoV-2 which was used to decipher three dominant epitopes based on the sera of 19 COVID-19 patients out of which two corresponded to the proteolytic cleavage sites S1 and S2 [51]. Thus, monitoring the host response to SARS-CoV-2 infection using custom microarrays has led to significant conclusions that are crucial to the development of useful diagnostic and prognostic markers and therapeutic strategies. It is also vital to evaluate the host response upon the use of an attenuated viral strain or nonpathogenic strain of a pathogen for vaccination and/or a vector to characterize any immunogenic response in host's body. Genetic host response has been evaluated in response to the most commonly used viral vectors; it was discovered that all adeno-associated viral vectors elicit a similar host response to the wild-type adeno-associated virus, whereas lentivirus vector did not when compared with wild-type counterparts [52]. The evaluation of non-primate host response to baculoviral vectors using cDNA microarray was substantial in producing potential strategies to reduce the immune response generated after intracranial injections [53]. The host's humoral response to a vaccine can also be evaluated using protein microarrays containing all or major proteins (antigens) of the virus against which the host has been vaccinated [54]. Other components of the host's response, for instance, the defense proteins (complement proteins, CD20 receptors, etc.), have also been evaluated upon vaccination for viral hemorrhagic septicemia virus in Japanese flounder [55].

Since these pathogenic viruses possess a great deal of replicating ability inside a host cell, they have been used as vectors for therapeutic purposes in their attenuated states. They are used to deliver a genetic segment that is intended to fix in the host genome to make up for the missing gene or to provide a genetic material of other microbes as vaccine vehicles, for instance, adenovirus-5. It is thus necessary to make sure of their safety along with their given efficacy. Host DNA microarrays have been used to differentiate between the genetic expression these attenuated vectors induce with that of the pathogenic counterparts to elucidate the safety of these vectors. Similarly, evaluating the genetic expression of the host in response to these

attenuated viruses is crucial as they are also used as live attenuated vaccines. For instance, a study has evaluated changes in the gene expression of cultured HeLa cells in response to attenuated modified vaccinia virus Ankara using a DNA microarray consisting of 15,000 human cDNAs [56]. The study revealed that the attenuated modified virus differentially regulated 11 genes which are involved in immune modulation and host resistance. Once a pathogen infects a host, it too must undergo changes that can be described either by evaluating its transcriptome or proteome. The pathogen side of the story of a host-pathogen interaction can also be unraveled by putting microarrays to use. The most basic concept is to isolate and compare the transcriptome or proteome of the pathogen that has been allowed to infect the host cell with a naïve pathogen [57]. However, this concept is still gaining attention for the analysis of change in viral genes and proteins after coming in contact with the host cell.

8.4 Conclusion

Over the years, the advancements in microarray technology have led to their evolution as a standard research tool of basic sciences. With their advent, it has become possible to conduct hybridization experiments for thousands of genes simultaneously. They have been successfully employed for global and specific expression profiling of different organisms. The ability to customize the microarrays has broadened their area of applicability from basic profiling to diagnostics and therapeutics in various diseased conditions. An interaction between a pathogen and its host is a two-way conversation which we can now decode with the intervention of microarrays and witness the revolution in pathogenesis, diagnostics, and therapeutics of infectious diseases brought about by it. Although challenging, this technology has the potential to identify patient-specific drugs having fewer side effects.

Acknowledgment HC would like to acknowledge the Council of Scientific and Industrial Research, Government of India, for providing financial assistance.

References

1. Govindarajan R, Duraiyan J, Kaliyappan K, Palanisamy M (2012) Microarray and its applications. *J Pharm Bioallied Sci* 4(Suppl 2):S310
2. Lemuth K, Rupp S (2015) Microarrays as research tools and diagnostic devices. In: *RNA and DNA diagnostics*. Springer, Cham, pp 259–280
3. Stears RL, Martinsky T, Schena M (2003) Trends in microarray analysis. *Nat Med* 9(1):140–145
4. Mello-Coelho VD, Hess KL (2005) A conceptual and practical overview of cDNA microarray technology: implications for basic and clinical sciences. *Braz J Med Biol Res* 38(10):1543–1552
5. Hurd PJ, Nelson CJ (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief Funct Genomic Proteomic* 8(3):174–183

6. Grunstein M, Hogness DS (1975) Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. *Proc Natl Acad Sci USA* 72(10):3961–3965
7. Bumgarner R (2013) Overview of DNA microarrays: types, applications, and their future. *Curr Protoc Mol Biol* 101(1):22-1
8. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270(5235):467–470
9. DeRisi J, Penland L, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM (1996) Use of a cDNA microarray to analyse gene expression. *Nat Genet* 14:457–460
10. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* 91(11):5022–5026
11. Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J, McCormick M (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res* 12(11):1749–1755
12. Blanchard AP, Kaiser RJ, Hood LE (1996) High-density oligonucleotide arrays. *Biosens Bioelectron* 11(6–7):687–690
13. Ferguson JA, Steemers FJ, Walt DR (2000) High-density fiber-optic DNA random microsphere array. *Anal Chem* 72(22):5618–5624
14. Nolan JP, Sklar LA (2002) Suspension array technology: evolution of the flat-array paradigm. *Trends Biotechnol* 20(1):9–12
15. Sosnowski RG, Tu E, Butler WF, O’Connell JP, Heller MJ (1997) Rapid determination of single base mismatch mutations in DNA hybrids by direct electric field control. *Proc Natl Acad Sci USA* 94(4):1119–1123
16. Heller MJ (2002) DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng* 4(1):129–153
17. Kricka LJ, Fortina P (2001) Microarray technology and applications: an all-language literature survey including books and patents. *Clin Chem* 47(8):1479–1482
18. Ochman H, Jones IB (2000) Evolutionary dynamics of full genome content in *Escherichia coli*. *EMBO J* 19(24):6637–6643
19. Preuss TM, Cáceres M, Oldham MC, Geschwind DH (2004) Human brain evolution: insights from microarrays. *Nat Rev Genet* 5(11):850–860
20. Shiu SH, Borevitz JO (2008) The next generation of microarray research: applications in evolutionary and ecological genomics. *Heredity* 100(2):141–149
21. Gerhold DL, Jensen RV, Gullans SR (2002) Better therapeutics through microarrays. *Nat Genet* 32(4):547–552
22. Kumar A, Pandey SC, Samant M (2020) DNA-based microarray studies in visceral leishmaniasis: identification of biomarkers for diagnostic, prognostic and drug target for treatment. *Acta Tropica* 208:105512
23. Chen R, Huang LH, Gao YY, Yang JZ, Wang Y (2019) Identification of differentially expressed genes in MG63 osteosarcoma cells with drug-resistance by microarray analysis. *Mol Med Rep* 19(3):1571–1580
24. Gmuender H (2002) Perspectives and challenges for DNA microarrays in drug discovery and development. *Biotechniques* 32(1):152–158
25. Fang Y, Lahiri J, Picard L (2003) G protein-coupled receptor microarrays for drug discovery. *Drug Discov Today* 8(16):755–761
26. Müller MB, Tang YW (2009) Basic concepts of microarrays and potential applications in clinical microbiology. *Clin Microbiol Rev* 22(4):611–633
27. Yoo SM, Keum KC, Yoo SY, Choi JY, Chang KH, Yoo NC, Yoo WM, Kim JM, Lee D, Lee SY (2004) Development of DNA microarray for pathogen detection. *Biotechnol Bioprocess Eng* 9(2):93–99
28. Wilson WJ, Strout CL, DeSantis TZ, Stilwell JL, Carrano AV, Andersen GL (2002) Sequence-specific identification of 18 pathogenic microorganisms using microarray technology. *Mol Cell Probes* 16(2):119–127

29. Hou Y, Zhang X, Hou X, Wu R, Wang Y, He X, Wang L, Wang Z (2018) Rapid pathogen identification using a novel microarray-based assay with purulent meningitis in cerebrospinal fluid. *Sci Rep* 8(1):1–10
30. Boriskin YS, Rice PS, Stabler RA, Hinds J, Al-Ghusein H, Vass K, Butcher PD (2004) DNA microarrays for virus detection in cases of central nervous system infection. *J Clin Microbiol* 42(12):5811–5818
31. Garaizar J, Rementeria A, Porwollik S (2006) DNA microarray technology: a new tool for the epidemiological typing of bacterial pathogens? *FEMS Immunol Med Microbiol* 47(2):178–189
32. Li J, Chen S, Evans DH (2001) Typing and subtyping influenza virus using DNA microarrays and multiplex reverse transcriptase PCR. *J Clin Microbiol* 39(2):696–704
33. Fitzgerald JR, Sturdevant DE, Mackie SM, Gill SR, Musser JM (2001) Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc Natl Acad Sci USA* 98(15):8821–8826
34. Zin NM, Al-Shaibani MM, Jalil J, Sukri A, Al-Maleki AR, Sidik NM (2020) Profiling of gene expression in methicillin-resistant *Staphylococcus aureus* in response to cyclo-(L-Val-L-Pro) and chloramphenicol isolated from *Streptomyces* sp., SUK 25 reveals gene downregulation in multiple biological targets. *Arch Microbiol*. <https://doi.org/10.1007/s00203-020-01896-x>
35. Frye JG, Lindsey RL, Rondeau G, Porwollik S, Long F, McClelland M, Jackson CR, Englen MD, Meinersmann RJ, Berrang ME, Davis JA (2010) Development of a DNA microarray to detect antimicrobial resistance genes identified in the National Center for Biotechnology Information database. *Microb Drug Resist* 16(1):9–19
36. Wilson M, DeRisi J, Kristensen HH, Imboden P, Rane S, Brown PO, Schoolnik GK (1999) Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization. *Proc Natl Acad Sci USA* 96(22):12833–12838
37. Kaveh R, Li YS, Ranjbar S, Tehrani R, Brueck CL, Van Aken B (2013) Changes in *Arabidopsis thaliana* gene expression in response to silver nanoparticles and silver ions. *Environ Sci Technol* 47(18):10637–10644
38. Smoot LM, Smoot JC, Graham MR, Somerville GA, Sturdevant DE, Migliaccio CAL, Sylva GL, Musser JM (2001) Global differential gene expression in response to growth temperature alteration in group A *Streptococcus*. *Proc Natl Acad Sci USA* 98(18):10416–10421
39. Butcher PD (2004) Microarrays for *Mycobacterium tuberculosis*. *Tuberculosis* 84(3–4):131–137
40. Steller S, Angenendt P, Cahill DJ, Heuberger S, Lehrach H, Kreutzberger J (2005) Bacterial protein microarrays for identification of new potential diagnostic markers for *Neisseria meningitidis* infections. *Proteomics* 5(8):2048–2055
41. Kafsack BF, Painter HJ, Llinás M (2012) New Agilent platform DNA microarrays for transcriptome analysis of *Plasmodium falciparum* and *Plasmodium berghei* for the malaria research community. *Malar J* 11(1):1–9
42. Wen Y, Ma WL, Li L, Wu QH, Xu QL, Zhang HY, Zheng WL (2005) Oligonucleotide microarray for human immunodeficiency virus detection. *Di I junyi da xuexuebao (Acad J First Med Coll PLA)* 25(3):293–297
43. De Assis RR, Jain A, Nakajima R, Jasinskas A, Felgner J, Obiero JM, Adenaiye O, Tai S, Hong F, Norris P, Stone M (2020) Analysis of SARS-CoV-2 Antibodies in COVID-19 convalescent plasma using a coronavirus antigen microarray. *BioRxiv*. <https://doi.org/10.1101/2020.04.15.043364>
44. Wang H, Wu X, Zhang X, Hou X, Liang T, Wang D, Teng F, Dai J, Duan H, Guo S, Li Y (2020) SARS-CoV-2 proteome microarray for mapping COVID-19 antibody interactions at amino acid resolution. *ACS Cent Sci* 6(12):2238–2249
45. Kato-Maeda M, Gao Q, Small PM (2001) Microarray analysis of pathogens and their interaction with hosts: technoreview. *Cell Microbiol* 3(11):713–719
46. Rubins KH, Hensley LE, Jahrling PB, Whitney AR, Geisbert TW, Huggins JW, Owen A, LeDuc JW, Brown PO, Relman DA (2004) The host response to smallpox: analysis of the gene

- expression program in peripheral blood cells in a nonhuman primate model. *Proc Natl Acad Sci USA* 101(42):15190–15195
47. Mehla R, Ayyavoo V (2012) Gene array studies in HIV-1 infection. *Curr HIV/AIDS Rep* 9 (1):34–43
 48. Judge M, Parker E, Nanche D, Le Souëf P (2020) Gene expression: the key to understanding HIV-1 infection? *Microbiol Mol Biol Rev* 84(2)
 49. Zhang X, Wu X, Wang D, Lu M, Hou X, Wang H, Liang T, Dai J, Duan H, Xu Y, Li Y (2020) Proteome-wide analysis of differentially-expressed SARS-CoV-2 antibodies in early COVID-19 infection. medRxiv. <https://doi.org/10.1101/2020.04.14.20064535>
 50. Jiang HW, Li Y, Zhang HN, Wang W, Yang X, Qi H, Li H, Men D, Zhou J, Tao SC (2020) SARS-CoV-2 proteome microarray for global profiling of COVID-19 specific IgG and IgM responses. *Nat Commun* 11(1):1–11
 51. Farrera-Soler L, Dagher JP, Barluenga S, Vadas O, Cohen P, Pagano S, Yerly S, Kaiser L, Vuilleumier N, Winssinger N (2020) Identification of immunodominant linear epitopes from SARS-CoV-2 patient plasma. *PLoS One* 15(9):e0238089
 52. Piersanti S, Martina Y, Cherubini G, Avitabile D, Saggio I (2004) Use of DNA microarrays to monitor host response to virus and virus-derived gene therapy vectors. *Am J Pharmacogenomics* 4(6):345–356
 53. Balasundaram G, Kwang TW, Wang S (2017) cDNA microarray assays to evaluate immune responses following intracranial injection of baculoviral vectors in non-human primates. *J Neurochem* 140(2):320–333
 54. Li B, Jiang L, Song Q, Yang J, Chen Z, Guo Z, Zhou D, Du Z, Song Y, Wang J, Wang H (2005) Protein microarray for profiling antibody responses to *Yersinia pestis* live vaccine. *Infect Immun* 73(6):3734–3739
 55. Byon JY, Ohira T, Hirono I, Aoki T (2006) Comparative immune responses in Japanese flounder, *Paralichthys olivaceus* after vaccination with viral hemorrhagic septicemia virus (VHSV) recombinant glycoprotein and DNA vaccine using a microarray analysis. *Vaccine* 24 (7):921–930
 56. Guerra S, López-Fernández LA, Conde R, Pascual-Montano A, Harshman K, Esteban M (2004) Microarray analysis reveals characteristic changes of host cell gene expression in response to attenuated modified vaccinia virus Ankara infection of human HeLa cells. *J Virol* 78 (11):5820–5834
 57. Rappuoli R (2000) Pushing the limits of cellular microbiology: microarrays to study bacteria–host cell intimate contacts. *Proc Natl Acad Sci USA* 97(25):13467–13469



Transcriptional Approach in the Identification of Drug Targets in *Candida* spp.

9

Mahnoor Patel, M. Amin-ul Mannan, and Banhishikha Datta

Abstract

High-throughput sequencing technologies have become essential in studies on genomics, epigenomics, and transcriptomics. While sequencing information has traditionally been elucidated using a low-throughput technique called Sanger sequencing, high-throughput sequencing (HTS) technologies are capable of sequencing multiple DNA molecules in parallel, enabling hundreds of millions of DNA molecules to be sequenced at a time. This advantage allows HTS to be used to create large data sets, generating more comprehensive insights into the cellular genomic and transcriptomic signatures of various diseases and developmental stages of disease-causing pathogens. The transcriptomics techniques like microarray and RNA sequencing (RNA-seq) can be used to compare differential expression of the genes and the underlying mechanism and regulatory pathways over diseased and normal states. In this chapter, we have elucidated the transcriptomics approach for the identification of the lead compounds for the diseases caused by *Candida* species. *Candida* spp. are commensal organisms and regarded as opportunist pathogens. It causes serious systemic infection with a mortality rate of ~50% in immunocompromised patients. The new clinical isolates are showing resistance to the existing drugs, and hence, new candidate molecules are required. The chapter enumerates the various technologies which can be deployed to identify the candidate drug molecules.

M. Patel · M. Amin-ul Mannan (✉) · B. Datta

Department of Molecular Biology & Genetic Engineering, School of Bioengineering & Biosciences, Lovely Professional University, Phagwara, Punjab, India
e-mail: mohammad.20597@lpu.co.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*,
https://doi.org/10.1007/978-981-16-0691-5_9

139

Keywords

Big data analysis · Next-generation sequencing · Integrated genomics · *Candida* spp. · Antifungals

9.1 Introduction

High-throughput omics approaches like genomics, transcriptomics, proteomics, and metabolomics contribute a great deal in understanding the biological process including the identification of candidate molecules for therapy. It can generate a large number of data in a single day [1]. However, all of these omics technologies face challenges like cleaning of data, identification of biomolecules, reduction of data dimensionality, biological contextualization, statistical validation, handling and storage of data, sharing, and archiving. Large-scale omics data set access is important for biological processes improvement and in systems biology. Since the procedural costs to experiment with high-throughput sequencing are far more economical as compared to costs a decade ago, it generates enormous data sets. On one hand, it is challenging, but at the same, it also creates an exhilarating opportunity for the biologists, biostatisticians, and computational biologists to analyze those data [2]. Omics approaches based on the global analysis of biological samples with an aid of high-throughput analysis and bioinformatics provide novel insights into the biological processes [1].

This chapter is an attempt to comprehend one of the omics approaches, i.e., transcriptomics, for the identification of the candidate molecules for the therapy of fungal disease caused by *Candida* spp. Candidiasis is an opportunistic fungal infection caused by *Candida* spp. Recent trends suggest that the number of cases and deaths related to candidiasis is alarming and escalating. Antifungal and multidrug resistance is one of the major challenges in the management of candidiasis. Among all *Candida* spp., *Candida auris* has emerged as a multidrug-resistant strain [3, 53]. The Centers for Disease Control and Prevention (CDC) has recognized it as a global threat. The transmission of this fungus resembles methicillin-resistant *Staphylococcus aureus* [4]. The chapter revolves around describing the omics approach for the identification of lead compounds and thereof its therapeutics.

9.2 Omics Approach

After the introduction of omics technologies in the post-genomic era, biological studies are characterized wisely and rapidly developed [5]. These kinds of technologies include genomics, proteomics, metabolomics, transcriptomics, lipidomics, and phenomics [52]. The omics approach is generally based upon the global analyses of biological samples with the help of high-throughput technology including bioinformatics which provides novel insights of the biological samples and their phenomena [6]. HTS technologies, like whole-exome sequencing, can be

used to identify novel variants and other mutations that may underlie many genetic disorders. The current high-density arrays with multiplexed features permit a sample size of ~20,000 cells with automated features and permit high-sample handlings. Numerous research and clinical applications like pharmaceuticals, diagnostics, therapeutics, disease prevention and pharmacogenomics, evolutionary genetics, and developmental biology including comparative genomics use the latest approach of genomics, proteomics, transcriptomics, and metabolomics [7]. Some of the tools used in drug discovery are mentioned in Table 9.1. A major fundamental difference between transcriptomics with other omics techniques is the activity measurement of a single class of molecules. The traditional methods require a different assay to measure the gene function, mutational analysis, metabolite and enzyme activity, and ligand-receptor interaction. Transcriptomics bridge the gap between genomics and proteomics and can aid in new drug discovery of economical, affordable, and better-quality drugs.

9.2.1 Genomics

Genomics is nucleic acid-based technology that relies upon several steps, namely, sample collection, high-quality extraction of nucleic acid, preparation of library, clonal amplification, and sequencing. Every approach is based on the downstream application of the sample. After sequencing, the process workflow includes cleaning of data, filtering, assembling, alignment, variant calling, annotation, and functional prediction [1, 2].

In every area of biological investigation, genomic technology is widely used. It includes genomics research which consists of functional as well as structural genomics. Three-dimensional structures of proteins that are encoded by a genome are also included in the structural genomics study [8]. It allows high-throughput methods for the structural analysis with the help of experimental and modeling approach combination. Today, the major branch of genomics is involved in sequencing the genome of the various organisms [9]. Describing genes and functions of proteins and their interaction with other proteins falls under preview of functional genomics. Bioinformatics and microarrays are significant tools for genomics. It includes metagenomics, epigenomics, and pharmacogenomics [1, 2, 10]. The genomics data of *Candida* spp. can be obtained from the *Candida* Genome Database which is maintained by the US National Institutes of Health [11]. The database has the represented genomes of *C. albicans*, *C. auris*, *C. dubliniensis*, *C. parapsilosis*, and *C. tropicalis*. Genomic sequences, including gene expression and protein information, can be retrieved from the database.

Table 9.1 Representative ‘omics’ approach used in drug discovery

Omics approach	Uses	Resource	Resource link
Genomics	<ul style="list-style-type: none"> • Mechanism of pathogenesis • Identification of virulent genes • Discovery of candidate molecules • Efficacy and toxicity of drugs 	• GWAS Central	https://www.gwascentral.org/
		• PharmGKB	https://www.pharmgkb.org/
		• dbGaP	https://dbgap.ncbi.nlm.nih.gov/http://www.candidagenome.org/
Proteomics	<ul style="list-style-type: none"> • Drug target efficacy • Protein toxicology • Protein-protein network interaction • Mass spectral database 	• ProteomicsDB	https://www.proteomicsdb.org/
		• The Human Protein Atlas	https://www.proteinatlas.org/
		• DITOP	http://bioinf.xmu.edu.cn/index.jsp https://massbank.eu/MassBank/
Transcriptomics	<ul style="list-style-type: none"> • High-throughput functional genomic data • Gene expression data linked to phenotype data • Minimum Information About a Microarray Experiment • Minimum Information About a High-throughput Nucleotide Sequencing Experiment • Understanding of cell pathways 	• GEO	https://www.ncbi.nlm.nih.gov/geo/
		• Open TG-GATES	https://toxico.nibiohn.go.jp/english/index.html
		• MIAME	https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html
		• MINSEQE	http://fged.org/projects/minseqe/
		• The LINCS Consortium	https://lincsproject.org/
Metabolomics	<ul style="list-style-type: none"> • Small-molecule metabolites found in the human body • GC/MS profiling studies of metabolites • Drug target efficacy and safety evaluation • Metabolic toxicity 	• Human Metabolome	https://hmdb.ca/
		• Golm Metabolome	http://gmd.mpimp-golm.mpg.de/
		• MetabolomeExpress	https://www.metabolome-express.org/

9.2.2 Proteomics

The proteome is the whole set of proteins of any organism which is translated by every organism or any biological system. Proteins differ in the presence of genetic and environmental changes. Proteomics comprises a particular type of cells, body

fluids, and tissues including their functions and structures. Proteomics is very helpful in understanding the research at the translation level after the genomics; it also helps in understanding the post-translation modifications. It is an essential tool for understanding the genome at its expression level [1, 2].

For the quantification of proteins from the multiple samples, proteomics approach is generally used. It uses both shotgun and targeted approaches. Recent advancement in mass spectrometer (MS) radically increases the sensitivity and decreases the sample amount which is required in the high-throughput analysis. It also allows for minimal differences in protein profusions and posttranslational modification identification [12]. The major steps in proteomics include proper sample collection, extraction of protein and peptides, enzymatic digestions, and fractionation/separation by using liquid chromatography followed by MS, identification, and quantification of proteins and peptides. Proteomics moved from the traditional 2D-PAGE-based spot extraction of proteins followed by the LC-MS or matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) system [1, 2, 10].

9.2.3 Transcriptomics

All the set of RNA molecules are coming under the transcriptomics studies are known as transcriptome. It includes mRNA, tRNA, and rRNA molecules with the other noncoding RNA molecules present in the cells. Unlike the whole genome, it varies under the influence of external conditions of the environment. It examines changes that occur in the entire transcriptome under different biological surroundings [13]. The various transcriptomics technologies are shown in Fig. 9.1.

RNAs are the sequence which is generated from the DNA sequences, and that is why they are the mirrors of DNA sequences. In the transcription process, RNA synthesis is the initial step of expression of the gene. Although same genome exists in every cell of an organisms, every cell expresses different genes at different transcriptional control to generate the diverse repertoire of proteins [14]. Transcriptomics data helps researchers in the understanding of gene function and its comparison of different types of healthy cells transcriptome to the transcriptome of the diseased cell. These type of data help researchers in understanding the genes' misleading functions and its interpretation [1, 2].

9.2.4 Metabolomics

Complex biochemical cascades of end products are generally known as metabolites. It can link with the genome, transcriptome, and proteome to a phenotype which can provide an important detail for the metabolic variation and complement in the genetic basis of discovery [15]. Metabolomics is generally used for the determination of the absolute and relative amount of sugars, amino acids, lipids, nucleotides,

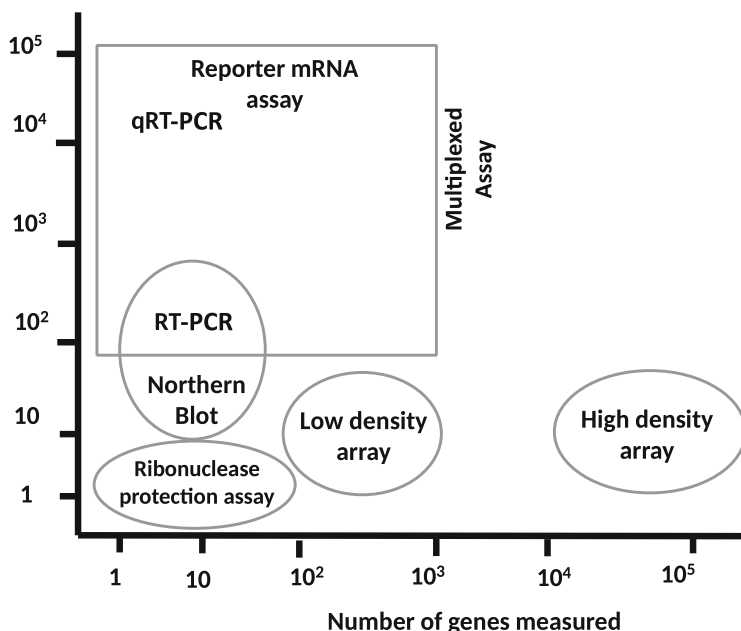


Fig. 9.1 Gene expression “transcriptomics” technologies (*the diagram is not drawn to the scale*)

steroids, and drug constituents [54]. Depending on instrumentation and applications in research, metabolomics can capture information of small molecules in liquid, solid, and capillary electrophoresis. Important steps in metabolomics research include experimental design, appropriate sample collection strategies, quenching of metabolism, extraction of optimized metabolites and its reconstitution from the sample, and data analysis of MS or NMR which includes data alignment, filtration, imputation, statistical analysis, annotation, and network or pathway analysis. These steps are extremely variable and depend on the sample analysis and platform used for the process [10, 16].

9.2.5 Pharmacogenomics

Every person possesses unique variants of the genome, which leads to an individual’s diverse reactions to drugs. Pharmacogenomics gives an idea about how a person’s genes affect drug responses so that safe and effective medications can be developed and the determination of its doses can be done. It helps in the discovery process of genes responsible for particular diseases including its investigation on the effects of genetic factors for medication for predicting a person’s response [17]. Confrontational drug reactions are described as an important cause for hospitalizations including deaths in few countries. Pharmacogenomics empowers researchers in understanding how inherited gene variances affect responses toward

medications. Pharmacogenomics gives opportunities to researchers for understanding the mechanism of differences of inherited disease-related genes affecting the patient's body response to the medications, in which data is important for a prediction about the drug's effectiveness and its response in the patient [1, 2].

9.2.6 Epigenomics

DNA sequence changes leads to the change in the gene expression, which are often heritable. Epigenetics studies the inheritance pattern in the gene expression without any changes in the DNA sequence per se. Epigenomics deals with the analysis of epigenetics at the global level and in the entire genome and genetic information in terms of DNA sequences as it is also able to affect the functions of particular genes [18]. There are five diverse mechanisms of epigenetic regulation: (1) methylation of DNA, (2) posttranslational modification of histone, (3) variants of histones, (4) RNA interference, and (5) nuclear organization. Genome function can be changed under the influence of exogenous factors which usually occurs in CpG islands, which is a GC-rich region of DNA based on methylation which is the most common genomic parameter (e.g., regions of the promoter, regulatory domains of genes, and intergenic regions of a genome) [1, 2].

9.2.7 Immunomics

Immunomics is the study of the regulation and response processes of the immune system against the pathogen. Immunomics deals with every molecule of the immune system including the targets of immune cells and their functions. There are many techniques related to genomics, bioinformatics, and proteomics which include immunomics [19]. After the advancement in genomics and proteomics, the immunomics approach uses bioinformatics, structural biology, high-throughput screening, and biochip for studying immune cells and their responses. Immunomics is generally used for discovering new susceptibility of genes and their correlation with the immune cells [20]. Every person's immune system possesses a great level of diversity as compared to the person's other body systems. For research on a highly complex system, traditionally developed methods are mostly limited. Immunomics may prove as an advanced newly developed approach. Generally, it is used for vaccine development including target identification and diagnosis of disease [1, 2, 51]. Immunological research became more effective with the help of the immunoinformatics approach which is also known as computational immunology. It applies in silico modeling including analyzing the problems and data of the immune system. This new branch of bioinformatics having several software and resources focused on immunology, helps in understanding complete immune system properties [21].

9.2.8 Cytomics

Cytomics comprises structural and functional research of the cellular systems. This kind of study involves databases at the genomic level. Cytomics studies also involve the use of many technologies at the genomic and proteomic levels [22]. Noninvasive, sensitive, fluorescence-based technologies are mostly involved in the studies of cytomics for conducting single-cell integrated analysis [23, 24]. Cell imaging and quantitative data of fluorescent technique which is performed on a single cell is helpful in a comprehensive analysis of cellular processes. Cytomics comprises current technologies including flow cytometry, confocal laser scanning microscopy, high-content screening, laser capture microdissection, bio-imaging, and laser scanning cytometry [25]. It provides approaches and strategies for the pharmaceutical research like a validation of target, development of drug, toxicological and pharmaceutical evaluation and validation, and efficiency at the clinical level for prognostic and personalized medicine [26].

9.3 Omics Application in Pharmaceutical Research

There are a few limitations in the application of the omics approach till now. Such omics data may create false-negative or false-positive results because of the large number of massive complicated data. Because of this limited accuracy and sensitivity of the methods, many times, few important functional biological molecules which are present in trace amount cannot be detected [27, 28]. Furthermore, the assessment of omics data lacks proper specificity. Nowadays, research in the pharmaceutical industry generally relies on the omics approach which includes genomics, proteomics, transcriptomics, and metabolomics. It also uses multiple combinations of omics technologies [29].

Every phase of pharmaceutical research which includes drug development, evaluation of efficacy, validation of target and discovery, safety assessment, and development of personalized medicine uses many kinds of omics approach. It is the most powerful and efficient tool in pharmaceutical discovery. It is becoming the most vital part of the network and systems biology which makes it possible for understanding in-depth concepts easily including the pathological processes and simulation performance of pathogen's interactions with the host immune system for all diseases [30]. The omics approach reveals all the possible key pathways and their mechanism for enhancing pharmaceutical research and drug development. Additionally, studying omics also highlights all the probable targets for new drug development, which allows safety assessment efficiently and personalized medicine development [31, 32].

9.3.1 Target Discovery

Target discovery is very essential for developing new drugs. In the past, new drug development process for any disease was dependent on the 500 early known drug

targets. After completion of the Human Genome Project back in 2003, the genomic studies indicate that there roughly 22,000 protein-encoding genes [33]. Till now, around 10% of genes have been explored for drug target identification, and still, there is much needed to be done. Generally, developing a drug based on a single chemical with a single target group is not efficient. In recent years, omics technologies with systems biology applied widely also provide an idea about the identification of target and novel drug development. In present times, there are many new omics technologies applied for designing new drugs including the discovery of targets, microbial genomics and proteomics, nuclear magnetic resonance, RNAi, gene transfection, and gene knockout modeling. These omics approach produces a vast number of data and many databases which have been constructed, like Online Mendelian Inheritance in Man (OMIM), Therapeutic Target Database (TTD), Cancer Gene Census, and Gene Expression Omnibus (GEO) [34, 35, 55] (Fig. 9.2).

9.3.2 Toxicity and Toxicogenomics

The toxicology of drugs plays an essential role in drug development and pharmaceutical research. Toxicity is the greatest reason for the terminating process of drug development. Toxicology of drugs can guide the clinical medication for reducing adverse reactions of drugs. From the last 20 years, many more omics technologies are applied for toxicology analysis in drug development and also promoted its discovery in the different fields of research in toxicology [36, 37].

The genomics application in the toxicology field is known as toxicogenomics. For clarifying the relationship between the changes in gene expression and toxicity, a toxicogenomics study is applied. It also helps in identifying probable genetic toxicants, and after that, their mechanism of action is understood. For understanding toxicogenomics, the microarray is generally used. It is reflected that practically, all toxic reactions depend on changes that occur in profiles of gene expression [56]. As compared with the traditional drug toxicity research, the new field of toxicogenomics delivers a more comprehensive and sensitive platform for the safety assessment of drugs. Measuring the expression of a gene at a larger scale, the most sensitive and relevant changes in genetics can be found which can be used for risk management as biomarkers. For example, gene expression which is involved in the repair of DNA damage may be a genotoxicity sign.

Transient changes at the earliest expression of genes are related to the stress response of the body, whereas long-term changes in the profile of genes are related to chronic toxicity [38]. It may become an adaptive response in the body. For the determination of chronic toxicity, carcinogenicity, and the drug's secondary toxic effects, this technology is very much essential. Furthermore, in the early stages of the new drug development process, precisely expressed genes or proteins specific to toxicants also developed like a biomarker for understanding and predicting drugs' potential toxicity. This generally helps in a lead compound generation for producing and evaluating toxicity with high efficiency and sensitivity. This mode of toxicity evaluation mode delivers more valuable and relevant information about the toxicity

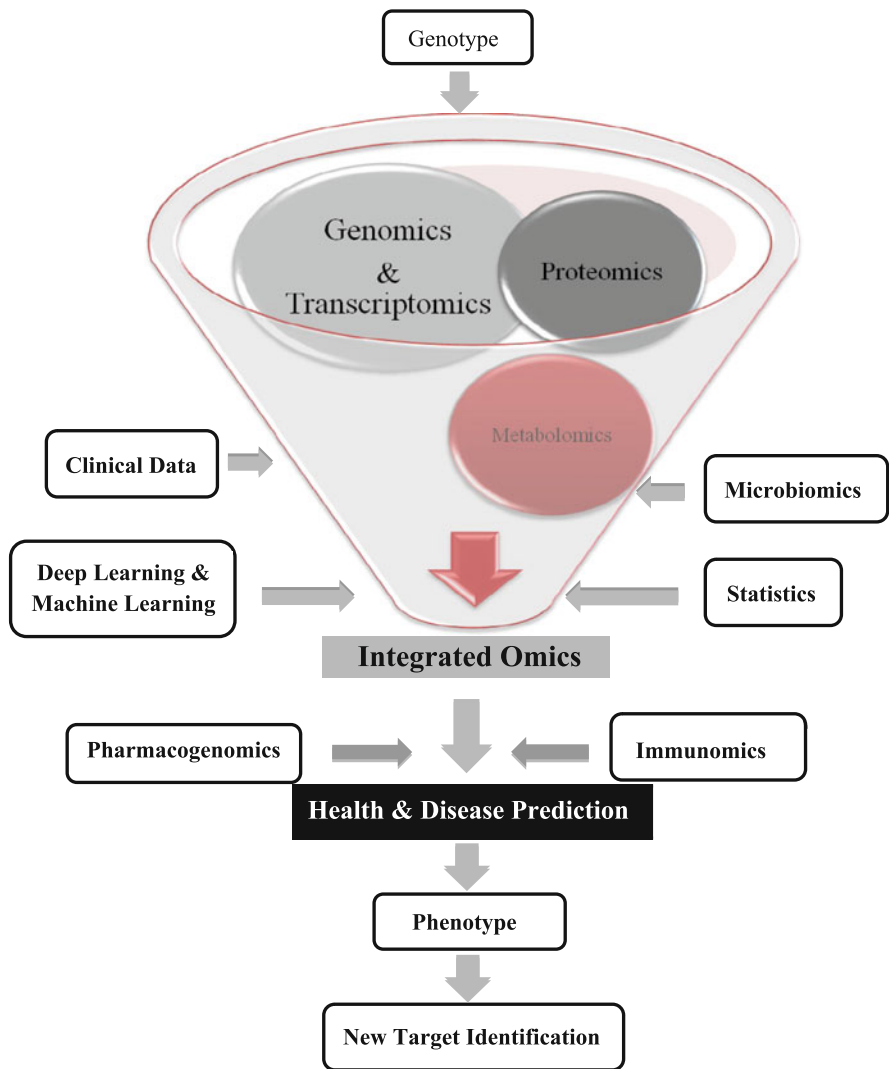


Fig. 9.2 The omics approach in the identification of new drug target

mechanism in a short time relatively. As compared to the toxicity study in a traditional manner, this newly developed omics technology which is known as toxicogenomics brings a revolution in drug toxicology studies [39].

9.3.3 Toxicoproteomics

As toxicogenomics is a larger field of study, toxicoproteomics is just a part of this vast field. Toxicoproteomics helps in identifying critical proteins and their pathways in biological systems that affected and respond to environmental exposure and adverse chemical reactions with the help of global expression technologies of proteins [40]. Traditional toxicology, expression analysis of differential proteins, and pathology are three major integrated areas of toxicoproteomics. Recently, this technology can reveal the expression of toxicant-reduced proteins; also, it can help study posttranslational modification with protein-protein interaction [41]. By doing a comparison of specific cells, organs, and tissues' protein expression profiles with those profiles which are generated by toxicants, toxicoproteomics can highlight in a very short period the specificity of the toxic protein expression which can execute functional molecule deficiency caused by toxicants. Consequently, with the help of the antibody analysis method, new markers of toxic proteins can also be discovered. These kinds of toxic markers can be applied for studying the mechanism of the human body at a safer dose [42].

9.4 Emerging Disease Causative Agents of *Candida* spp.

9.4.1 Prevalence of *Candida* Species

Candida spp. cause superficial skin infection to mucosal and deep tissue infections. It contributes to high mortality and morbidity [3]. The emergence of multidrug-resistant clinical isolates and a limited number of antifungal agents adds to the seriousness of the problem. One of the major problems with candidiasis or candidemia is biofilm formation [43]. Severely ill or immunocompromised patients are generally more prone to developing both superficial and life-threatening infections [44]. It is also a very common infection in AIDS patients which leads to malnutrition and causes interference in the absorption of the medications which was proved in epidemiological studies [43].

9.4.2 Prevalence of Non-albicans *Candida* Species

Generally, *Candida albicans* is the predominant species involved in invasive fungal infections. However, recent literature suggested there are increasing cases of emergence of non-albicans *Candida*. These epidemiological changes are associated with severe immunosuppression or critical illness and broad-spectrum antibiotic exposure with increasing age [44]. An investigation exhibited that more than half of the cases of infections caused by *Candida* species were reportedly by *Candida albicans*, and the other incidence rate for the non-albicans infection rates was reportedly 14% by *Candida parapsilosis* and *Candida glabrata* including 7% by the *Candida tropicalis* and 2% by *Candida krusei* [43, 45].

9.4.3 Targets of Antifungal Candidates

When compared to antibacterial research, slight advancements have occurred in the development of a new antifungal agent. As fungi are eukaryotic organism having a close evolutionary relationship with the human host, this creates complications in the search for the antifungal agents. The help of new approaches about antifungal therapy including target identification and rational drug design technologies provides imperative acceleration in the development process of an antifungal agent by reducing the time for the cure and improving the quality of patient's life. Nowadays, commercially available antifungal agents have targets that are restricted to the plasma membrane and cell wall. Some examples of antifungal candidate's new as well as old targets are mentioned in Table 9.2 [43, 57].

The evolution of drug resistance is a rapid phenomenon in *Candida* spp. Some of the online database tools that can be used for drug target identification are mentioned in Table 9.3. For developing a new drug, the first step is to identify the drug target and its validation. It is also very much essential for the elucidation of disease pathology mechanism identification and the effects of drugs [46]. Using the in silico approach, our lab also contributed to the identification of target molecules in multidrug-resistant *C. auris* (unpublished data). Even though these tools help in the identification of the drug molecule, however, it needs to be validated in wet lab experiments.

9.4.4 NGS and Fungal Diagnosis

The next generation sequencing offer valuable tools in understanding the molecular mechanisms of antifungals compounds. It can also be depolyed for the detection of new mycobiota and species specific identification. In mycological diagnosis and research, sequencing technology including an enhanced capability provides a

Table 9.2 Antifungal candidates with targets

Antifungal candidates	Target
Echinocandins	Inhibition of β -glucan synthesis
Nikkomycin and polyoxins	Inhibition of chitin synthase
Azoles	Inhibition of 14 α -demethylase
Terbinafine and naftifine	Inhibition of squalene epoxidase
Polyenes, naphthoquinones, eugenol analogues, isoquercitrin	Bind to ergosterol
5-Flucytosine	Inhibition of DNA synthesis
Amphotericin B, miconazole, ciclopirox	Production of reactive oxygen species (ROS), leading to cell death
Sordarins	Inhibition of protein synthesis
Griseofulvin	Microtubule assembly
Triphenylethylenes	Inhibition of calcineurin signaling

Table 9.3 Software/tool/databases used for new drug target identification

Software/tool/database	Purpose
UniProt	The whole proteome of <i>Candida</i> can be retrieved
BLASTP	Using this tool, the retrieved proteins can be compared with the Human Protein database, and their foreignness can be determined
PVS and EMBOSS	Using these tools, antigenicity of the predicted protein can be identified. The score can be analyze using EMBOSS, and an antigenic propensity graph can be generated using PVS
ArgusLab	Used for the generation of antigenic peptide model from antigenic protein including energy and geometry calculation for each peptide
Swiss-MODEL	Used for the 3D model generation of antigenic protein
ProtParam	Used for physicochemical properties analysis of antigenic protein
SOPMA	Secondary structure prediction is done by using SOPMA
PROCHECK	Used for stereochemical quality analysis of immunogenic protein
ProSA-web	Used for calculation of Z-score
Kolaskar and Tongaonkar Method	Used for categorizing linear and conformation B-cell epitopes
Emini Surface Accessibility	Used for analyzing surface accessibility of the protein
Karplus and Schulz Flexibility	Used for analyzing the flexibility of the protein
Parker Hydrophilicity	Used for analyzing hydrophilicity of the protein
ElliPro	Used for analyzing conformational B-cell epitope
NetMHCIIpan 4.0	Used for analyzing helper T cell from immunogenic protein
NetCTL 1.2	Used for analyzing cytotoxic T cell from immunogenic protein
IEDB Immunogenicity	Used for analyzing strong and weak immunogenicity of the predicted CTL
PEP-FOLD3	Used for the 3D model generation of predicted CTL molecule
PatchDock	Used for performing molecular docking studies between HLA-A*0201 molecules and predicted CTL molecule
FireDock	Used for refining molecular docking results
RasMol	Used for visualization of molecular docking results

powerful tool. Next-generation sequencing (NGS) functionality is applied in the public health microbiology laboratories for the studies of metagenomics and outbreak monitoring. Speed and sensitivity of diagnosis of infectious fungal diseases including determination of the mycobiome can be increased with the help of advancement in molecular tools and techniques. NGS can enhance the creation of data at the molecular level because the mycobiota which is dependent on culture for identification method is the major limitation of many fungal species that cannot culture in vitro, and at the same time, NGS technology is very much valuable in the diagnosis. As it is true that the capability of the sequencer is limited, it means an application of the whole-genome sequencing for complex microbiome determining and diagnosis in the medical sample is far beyond the possibility which exists in today's sequencers. Nowadays, the increasing capacity of the sequencing platform with the continuous decrease in sequencing cost makes it a striking tool in mycology

Table 9.4 Comparative identification methods for the *Candida* spp.

Sr. no.	Method	Advantages	Disadvantages
1	Microbial	Cost-effective	Time-consuming
2	Molecular	Highly accurate	High cost
3	MALDI-TOF	Rapid and accurate identification	Cost is high
4	VITEK	Automatic identification	Cost is high
5	NGS	Accuracy is high	Cost is high, can't afford by every lab

for microbiome analysis [47], for example, *Candida glabrata*, *Candida auris*, *Candida tropicalis*, *Candida albicans*, and *Candida parapsilosis* (Table 9.4).

9.5 Current Challenges with Future Directions

Due to rapid advancement in high-throughput technologies and computational data analysis, the omics technology is getting wider publicity and acceptance. The transcriptional analysis is rapidly implemented in drug discovery owing to its sensitivity, large-scale quantitative data, reproducibility, and robust assay method. In a single day with the automated microplate reader or RNA sequencing, trillions of data sets can be captured. However, handling large data requires a specialist or data scientist with a background in biology, and often, it is difficult to obtain [48]. After 20 years of exploration of omics technologies, it became routine work to generate and deal with omics data which is not so much tedious about the generation of data sets with the help of high-throughput data by an analytical approach.

In the nearest future, for describing biology processes, multiple omics technology combinations are the general approach for assessment. It produces a complex large number of data at various levels which include DNA, RNA, and proteins [49]. By experimenting or with the help of exploration of Internet databases, omics data can be acquired. But it is more difficult for processing because of many reasons, like data-type diversity, redundancy of database, and uniformity lacking in description of standard data. How to deal with this kind of a large amount of data especially taken from different sources of multi-omics approach is the most difficult challenge in omics research. Network biology may become a probable solution to this problem for its efficient solution, as it describes biochemical systems like a network system for multi-omics data [50].

Acknowledgment Dr. Mannan would like to acknowledge lab funding from DST-SERB for a core research grant (EMR/2017/002299) vide sanction order SERB/F/10997/2018-2019.

Funding The work was not directly funded from any research grant.

References

1. Mishra BB, Langefeld C, Olivier M, Cox LA (2020) Integrated Omics: tools, advances and future approaches. *J Mol Endocrinol* 62:21–45
2. Kai YS, Hui LR, Zi JH, Ru LX, Ji Y, Lei S, Dong ZW (2015) Omics in pharmaceutical research: overview, applications, challenges, and future perspectives. *Chin J Nat Med* 13:3–21
3. Park JY, Bradley N, Brooks S, Burney S, Wassner C (2019) Management of patients with *Candida auris* fungemia at the community hospital, Brooklyn, New York, USA, 2016–2018. *Emerg Infect Dis* 25:3
4. Torres SR, Kim HC, Leach L, Chaturvedi S, Bennett CJ, Hill DJ, Jesus MD (2019) Assessment of environmental and occupational exposure while working with multidrug-resistant (MDR) fungus *Candida auris* in an animal facility. *J Occup Environ Hyg* 14:7
5. Hume HKC, Vidigal J, Carrondo MJT, Middelberg APJ, Roldao A, Lua LHL (2019) Synthetic biology for bioengineering virus-like particle vaccines. *Biotechnol Bioeng* 116:919–935
6. Paananen J, Fortino V (2019) An omics perspective on drug target discovery platforms. *Brief Bioinform*:1–17
7. Doolan DL, Apte SH, Proietti C (2014) Genome-based vaccine design: the promise for malaria and other infectious diseases. *Int J Parasitol* 44:901–913
8. Bredel M, Jacoby E (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat Rev Genet* 5:262–275
9. Mishra NK, Shukla M (2014) Application of computational proteomics and lipidomics in drug discovery. *J Theor Comput Sci* 1:105
10. Wolfender JL, Litaudon M, Touboul D, Queiroz EF (2019) Innovative omics-based approaches for prioritization and targeted isolation of natural products – new strategies for drug discovery. *Nat Prod Rep* 36:855–868
11. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G (2017) The *Candida* Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res* 45(D1):D592–D596
12. Nami S, Mohammadi R, Vakili M, Khezripour K, Mirzaei H, Morovati H (2019) Fungal vaccines, mechanism of actions and immunology: a comprehensive review. *Biomed Pharmacother* 109:333–344
13. Ravinarayanan H, Coico R, Sundar K (2015) Identification of putative therapeutic targets in *Candida tropicalis*: an *in-silico* approach. *Trends Bioinform* 8:52–62
14. Medici NP, Poeta MD (2015) New insights on the development of fungal vaccines: from immunity to recent challenges. *Mem Inst Oswaldo Cruz* 110:966–973
15. Khan SR, Baghdasarian A, Fahlman RP, Michail K, Siraki AG (2014) Current status and prospects of toxicogenomics in drug discovery. *Drug Discov Today* 19:562–578
16. Jorge S, Dellagostin OA (2017) The development of veterinary vaccines: a review of traditional methods and modern biotechnology approaches. *Biotechnol Res Innov* 1:6–13
17. Perlin DS, Richardson RR, Izquierdo AA (2017) The global problem of antifungal resistance: prevalence, mechanisms, and management. *Lancet* 17(12):e383–e392
18. Mochon B, Cutler JE (2005) Is a vaccine needed against *Candida albicans*? *Med Mycol* 43:97–115
19. Cole ST (2002) Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *Eur Respir J* 20:78–86
20. Taki T (2013) Bio-recognition and functional lipidomics by glycosphingolipid transfer technology. *Proc Jpn Acad* 89(7):302–320
21. Raghuvanshi R, Singh M, Shukla V (2018) Immunoinformatic approaches in epitope prediction for vaccine designing against viral infections. *Virol Immunol J* 2:2
22. Redi D, Raffaelli CS, Rossetti B, Luca AD, Montagnani F (2018) *Staphylococcus aureus* vaccine preclinical and clinical development: current state of the art. *New Microbiol* 41:208–213

23. Csermely P, Korcsmaros T et al (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery A comprehensive review. *Pharmacol Ther* 138:333–408
24. Zhang W, Li F, Nie L (2010) Integrating multiple ‘omics’ analysis for microbial biology: application and methodologies. *Microbiology* 156:287–301
25. Kumar G, Chaudhary KK, Misra K, Tripathi A (2017) Next-generation sequencing for drug designing and development: an omics approach for cancer treatment. *Int J Pharm* 13:709–723
26. Wheelock CE, Goss VM, Balgoma D et al (2013) Application of omics technologies to biomarker discovery in inflammatory lung diseases. *Eur Respir J* 42:802–825
27. Chassey B, Meyniel-Schicklin L, Aublin-Gex A, Andre P, Lotteau V (2012) New horizons for antiviral drug discovery from virus-host protein interaction networks. *Curr Opin Virol* 2:606–613
28. Cui T, Zeng J, He ZG (2018) Anti-tuberculosis drug target discovery by targeting the higher in-degree proteins (HidPs) of the pathogen’s transcriptional network. *J Tuberc I*
29. Carvalho A, Duarte-Oliveira C et al (2017) Fungal vaccines and immunotherapeutics: current concepts and future challenges. *Curr Fungal Infect Rep* 11:16–24
30. Pais P, Galocha M et al (2019) Microevolution of the pathogenic yeasts *Candida glabrata* during antifungal therapy and host infection. *Microb Cell* 6:142–159
31. Bencurova E, Gupta SK, Sarukhanyan E, Dandekar T (2018) Identification of antifungal targets based on computer modelling. *J Fungi* 4:81
32. Li X, Hou Y, Yue L, Liu S, Du J, Sun S (2015) Potential targets for antifungal drug discovery based on growth and virulence in *Candida albicans*. *Antimicrob Agents Chemother* 59:5885–5891
33. Kumar S, Kushwaha PP, Gupta S (2019) Emerging targets in cancer drug resistance. *Cancer Drug Resist* 2:61–77
34. Bar E, Gladiator A et al (2020) A novel Th cell epitope of *Candida albicans* mediates protection from fungal infection. *J Immunol* 188:5636–5646
35. Lattif AA, Mukherjee P et al (2011) Lipidomics of *Candida albicans* biofilms reveals phase-dependent production of phospholipid molecular classes and role for lipid rafts in biofilm formation. *Microbiology* 157:3232–3242
36. Cowell AN, Winzeler EA (2019) Advances in omics-based methods to identify novel targets for malaria and other parasitic protozoan infections. *Genome Med* 11:63
37. Sexton AE, Doerig C, Creek DJ, Carvalho TG (2019) Post-genomic approaches to understanding malaria parasite biology: linking genes to biological functions. *Infect Dis* 5:1269–1278
38. Becker JM, Kauffman SJ et al (2010) Pathway analysis of *Candida albicans* survival and virulence determinants in a murine infection model. *PNAS* 107:22044–22049
39. Bagnoli F, Baudner B et al (2011) Designing the next generation of vaccines for global public health. *J Integr Biol* 15:545–566
40. Toth R, Cabral V et al (2018) Investigation of *Candida parapsilosis* virulence regulatory factors during host-pathogen interaction. *Sci Rep* 8:1346
41. Lazo JS, McQueeney KE, Sharlow ER (2017) New approaches to difficult drug targets: the phosphatase story. *SLAS Discov* 22:1071–1083
42. Van Vleet TR, Liguori MJ et al (2019) Screening strategies and methods for better off-target liability prediction and identification of small-molecule pharmaceuticals. *SLAS Discov* 24:1–24
43. Sardi JCO, Scorzoni L, Bernardi T, Fusco-Almeida AM, Mendes Giannini MJS (2013) *Candida* species: current epidemiology, pathogenicity, biofilm formation, natural antifungal products and new therapeutic options. *J Med Microbiol* 62:10–24
44. Cortegiani A, Misseri G, Fasciana T, Giammanco A, Giarratano A, Chowdhary A (2018) Epidemiology, clinical characteristics, resistance, and treatment of infections by *Candida auris*. *J Intensive Care* 6:69
45. Maheshwari M, Kaur R, Chadha S (2016) *Candida* species prevalence profile in HIV seropositive patients from a major tertiary care hospital in New Delhi, India. *J Pathog.* <https://doi.org/10.1155/2016/6204804>

46. Malule HR, Lopez-Agudelo VA, Gomez-Rois D (2020) *Candida auris*: a bibliometric analysis of the first ten years of research (2008-2018). *J Appl Pharm Sci* 10:12–21
47. Zoll J, Snelders E, Verweij PE, Melchers WJE (2016) Next-Generation sequencing in the mycology lab. *Curr Fungal Infect Rep* 10:37–42
48. Opathy C, Gabaldon T (2019) Recent trends in molecular diagnostics of yeast infections: from PCR to NGS. *FEMS Microbiol Rev* 43:517–547
49. Nandikolla SK, Shaik M, Varali S, Seelam R (2011) Emerging trends in various fields with systems biology approach. *J Comput Sci Syst Biol* 13. <https://doi.org/10.4172/0974-7230.S13-004>
50. Korcsmaros T, Szalay MS, Bode S, Kovacs IA, Csermely P (2007) How to design multi-target drugs: target search options in cellular networks. *Expert Opin Drug Discov* 2:1–10
51. Cotugno N, Ruggiero A et al (2019) OMIC technologies and vaccine development: from the identification of vulnerable individuals to the formulation of invulnerable vaccines. *J Immunol Res*. <https://doi.org/10.1155/2019/8732191>
52. Kandpal RP, Saviola B, Felton J (2009) The era of omics unlimited. *Biotechniques* 46:351–355
53. Parente-Rocha J, Bailao AM et al (2017) Antifungal resistance, metabolic routes as drug targets, and new antifungal agents: an overview of endemic dimorphic fungi. *Mediators Inflamm*. <https://doi.org/10.1155/2017/9870679>
54. Roessner U, Bowne J (2009) What is metabolomics all about? *Biotechniques* 36:363–365
55. Roti G, Stegmaier K (2012) Genetic and proteomic approaches to identify cancer drug targets. *Br J Cancer* 106:254–261
56. Subhashini R, Jeyam M (2017) Computational identification of putative drug targets in *Malassezia globosa* by subtractive genomics and protein cluster network approach. *Int J Pharm Pharm Sci* 9:215–221
57. Xiao G, Zhang X, Gao Q (2017) Bioinformatic approaches for fungal omics. *BioMed Res Int*. <https://doi.org/10.1155/2017/7270485>



Noncoding RNA Profiling: Potential Application in Infectious Diseases

10

Shiffali Khurana, Uma Dhawan, and Vibha Taneja

Abstract

A huge diversity of noncoding RNAs (ncRNAs), which lack the protein-coding ability, are transcribed from the mammalian genome. The ncRNAs act as important mediators for target gene expression by regulating the levels of transcription, translation, and degradation. Among the known ncRNAs, microRNAs (miRNAs) as well as long noncoding RNAs (lncRNAs) have been of key interest for various human pathologies. This chapter briefly summarizes the importance of microRNAs in the context of pathogenic viral infections. Getting insight into the role of host miRNAs during viral infections can provide clues toward better understanding and identification of novel therapeutic strategies against specific viral infections.

Keywords

MicroRNAs · DNA viruses · RNA viruses

S. Khurana

Department of Biomedical Science, Bhaskaracharya College of Applied Sciences, University of Delhi, New Delhi, India

Department of Research, Sir Ganga Ram Hospital Delhi, New Delhi, India

U. Dhawan

Department of Biomedical Science, Bhaskaracharya College of Applied Sciences, University of Delhi, New Delhi, India

V. Taneja (✉)

Department of Research, Sir Ganga Ram Hospital Delhi, New Delhi, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_10

157

10.1 Introduction

The discovery of a novel class of RNAs that do not code for proteins was a major breakthrough. These diverse non-translational RNA molecules are involved in posttranscriptional gene regulation. Depending upon their length, localization, and function, the noncoding RNAs can be classified as small and long noncoding RNAs (sncRNAs and lncRNAs, respectively). The sncRNAs include miRNAs, PIWI-interacting RNAs (piRNAs), and small nuclear and nucleolar RNAs (snRNAs and snoRNAs) [1].

Because of endogenous origin, miRNAs are considered of utmost importance for their role in gene regulation. miRNAs are small noncoding RNAs and generally comprise of ~21–22 nucleotides. After the identification of first miRNA, *lin-4* in *Caenorhabditis elegans* [2], miRNAs are extensively found in animals, plants, bacteria, and viruses. To date, ~2700 mature miRNAs (<http://www.mirbase.org>) have been described in humans and are known to regulate approximately 60% of coding RNA transcriptome [3, 4].

Deregulated expression of miRNAs is predominantly associated with several pathologies like cancer, neurodegenerative, cardiovascular, and immune disorders. Infectious diseases pose a great risk to human health, although tremendous effort has been made to elucidate the mechanism of pathogenesis and safeguard and cure these diseases. miRNAs are key regulators of infectious diseases caused by bacteria, viruses, parasites, and fungi. The first evidence for altered expression of miRNAs was described for viral infections [5]. Host-pathogen interactions result in altered expression of miRNAs in the host and have been implicated for various infectious diseases. miRNAs act as key players of host reaction to infection by modulating the proteins active in the immune system [6, 7]. The manipulation of host cellular miRNAs is a survival strategy for various bacteria and viruses and results in efficient replication and pathogenesis inside the host.

10.2 Biogenesis of miRNAs

The initiation of miRNA synthesis occurs in the nucleus and involves the transcription of miRNA genes by RNA polymerase II into a long double-stranded primary miRNA transcript (pri-miRNA) [8]. The steps for miRNA biogenesis take place in two compartments (i.e., nucleus and cytoplasm) and involve the action of distinct protein complexes [9]. The class 2 RNase III enzyme Drosha (Pasha in *Drosophila melanogaster*) recognizes and cleaves at the base of the stem of pri-miRNA to generate a hairpin structure, precursor miRNA (pre-miRNA), which consists of approximately 70 nucleotides with 2-nucleotide 3' overhang [10, 11]. Another gene, DiGeorge syndrome critical region gene 8 (DGCR8), makes a complex with pri-miRNA. DGCR8 specifically recognizes the junction between single-stranded (ss) and dsRNA regions of pri-miRNA stem and mediates cleavage of RNA duplex 11 bp away from this site via Drosha [12]. Exportin-5 (Exp5)/RanGTP complex exports the pre-miRNA out of the nucleus where Dicer (RNase III) and TRBP

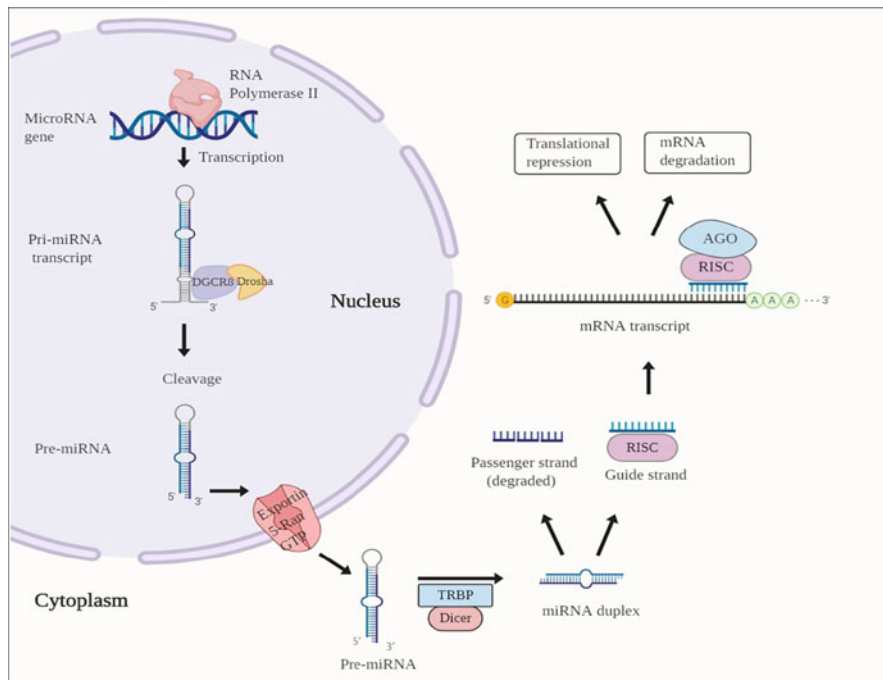


Fig. 10.1 Schematic representation of biogenesis of miRNAs. miRNA processing occurs in two compartments, i.e., nucleus where transcription of miRNA gene and cleavage of pri-miRNA take place and then transported to cytoplasm where pre-miRNA matures into miRNA duplex

(transactivation response element RNA-binding protein) complex binds and processes pre-miRNA into ~18- to 23-nucleotide-long mature miRNA duplex [13]. One of these duplex strands interacts with RNA-inducing silencing complex (RISC) and Argonaute (AGO) protein which functions as mature miRNA or “guide strand” and regulates the target gene expression in the cytoplasm. The second strand, known as “passenger strand,” is degraded by cellular machinery [14] (Fig. 10.1). The seed sequence (2–7 nucleotides) present at the 5′ end of guide strand binds with target mRNA. Binding sites for miRNAs are there in the 3′UTR, 5′UTR, coding, and promoter regions. Also, they have been identified to act as both activators and repressors of gene transcription [15–17].

10.3 Mechanisms of Action of miRNA

The guide strand interacts with AGO2 protein to generate RISC and become functional. The miRNA guides the RISC to target mRNA and mediates gene expression through (1) site-specific mRNA cleavage, (2) translational repression, or (3) mRNA degradation.

10.3.1 Site-Specific mRNA Cleavage

Argonaute proteins are evolutionary conserved proteins and form the core of RISC. They contain PAZ and PIWI domains as the main structural components. PAZ domain has a specific binding site for the 2-nucleotide 3' overhang of pre-miRNA [18, 19]. PIWI domain contains a conserved aspartate-aspartate-glutamate motif and shows similarity to RNase H [20, 21], implying that AGO protein holds mRNA cleaving “Slicer” activity. Of all known human AGO proteins, only AGO2 containing RISC promotes mRNA cleavage [22, 23].

10.3.2 Translation Repression

miRNA/RISC (1) competes with eIF4E (eukaryotic initiation factor 4E) and represses translational initiation at 5' cap region [24], (2) inhibits the circularization of mRNA, and (3) inhibits the assembly of 60S ribosome subunit with 40S preinitiation complex, further repressing translation [25]. Another study suggested that target mRNAs could accumulate in processing bodies (P bodies). P bodies do not have any components required for translation, and hence repression is achieved by relocalization of target mRNAs to P bodies [26].

10.3.3 mRNA Degradation

P bodies are also crucial for mRNA degradation. The sequestered mRNA can be degraded via deadenylation followed by decapping by Dcp1/2enzymes and Xrn1 exonuclease. Besides, mRNA degradation also requires the interaction of AGO protein and GW182 (RNA-binding protein) [27]. However, the exact mechanism of target mRNA degradation is still unknown.

10.4 MicroRNAs and Viral Infections

Multiple independent studies have analyzed the role of miRNAs in infectious diseases. Further, the chapter focuses on the modulation of host miRNAs as a response to infection caused by pathogenic DNA and RNA viruses and their role in immune regulation. In addition, the role of viral miRNAs is also highlighted.

10.4.1 DNA Viruses

miRNAs have been recognized as important players in DNA virus infections, modulating viral replication and immune responses and promoting cellular proliferation. We have highlighted the major effects produced by miRNAs in hepatitis B virus and herpesvirus in Fig. 10.2.

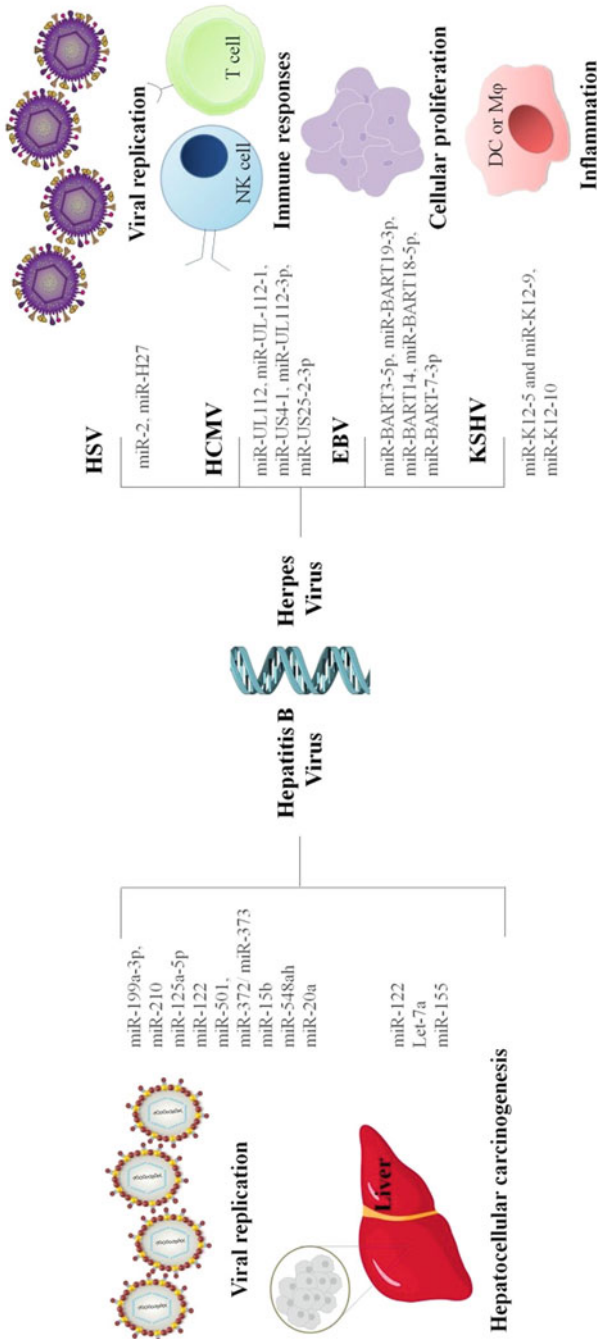


Fig. 10.2 DNA viruses and miRNAs. In hepatitis B virus and herpesvirus, certain miRNAs promote viral infection by modulating viral replication and immune responses and by promoting cellular proliferation and hepatocellular carcinoma

10.4.1.1 Hepatitis B Virus and miRNAs

Hepatitis B virus (HBV), a *Hepadnaviridae* virus, is a leading cause of liver disorders varying from acute to chronic hepatitis, cirrhosis of liver, and hepatocellular carcinoma (HCC).

Accumulating evidences suggest the role of cellular miRNAs in modulating hepatitis B virus replication. Several miRNAs either bind directly to the viral transcript or indirectly target host cellular factors implicated in viral replication and pathogenesis as tabulated in Table 10.1. For example, miR-199a-3p targets coding region, and miR-210 targets pre-S1 region of the surface antigen of HBV

Table 10.1 Summary of miRNAs implicated in HBV infection

miRNA	Expression after infection	Target gene (s)	Functions
miRNAs that directly or indirectly modulate HBV replication			
miR-199a-3p	Down	HBsAg coding	Reduces viral replication
miR-210	Down	HBsAg pre-S1 region	Reduces viral replication
miR-125a-5p	Up	HBsAg mRNA	Translational inhibition
miR-122	Down	Cyclin G1	Inhibits cellular proliferation
miR-501 and miR-372/miR-373	Up	HBxIP	Promotes HBV replication
miR-15b	Down	HNF1 α	Modulates HBV replication
miR-548ah	Up	HDAC4	Promotes HBV replication
miR-20a	Up	HBV cccDNA methylation	Suppresses HBV replication
miRNAs implicated in HBV-associated HCC and liver cirrhosis/fibrosis			
miR-122	Down	NDRG3, PTTG1	Enhanced proliferation of HCC cells
Let-7a	Down	STAT3	Promotes cellular proliferation and hence leads to HCC
miR-155	Up	SOCS1	Represses HBV progression by targeting SOCS1, ultimately activating JAK/STAT pathway
miR-19b	Down	TGF- β	Protects cells from fibrosis
miR-150 and miR-194	Down	C-myb and rac1, respectively	Inhibit ECM production

Up upregulated, *Down* downregulated, *HBsAg* HBV surface antigen, *HBxIP* HBx interacting protein, *HNF1 α* hepatocyte nuclear factor 1 α , *HDAC4* histone deacetylase 4, *NDRG3* N-myc downstream-regulated gene 3, *PTTG1* pituitary tumor-transforming gene 1, *STAT3* signal transducer and activator of transcription 3, *SOCS1* suppressor of cytokine signaling 1, TGF- β transforming growth factor- β

(HBsAg) [28], thereby reducing viral replication. This was the first report showing the role of miRNAs in regulating HBV replication. Another miRNA, miR-125a-5p inhibits translation by binding to HBsAg mRNA [29]. miR-122 impedes viral replication by upregulating the production of cellular heme oxygenase-1 (HO-1) that has been shown to result in lower levels of HBV cccDNA [30, 31]. Interestingly, a considerable decrease in miR-122 levels was observed in patients infected with HBV as compared to noninfected individuals. miR-122 specifically targeted cyclin G1, which interacts with p53, and hence the binding of p53 to HBV enhancer elements is obstructed which further inhibits HBV transcription [32]. Contrarily, the upregulation of miR-501 and miR-372/373 in the host promoted HBV replication by targeting HBx interacting protein (HBxIP) and nuclear factor 1B (NF-1B), respectively [33, 34]. Similarly, increased expression of HBx protein resulted in downregulation of miR-15b. miR-15b directly binds the hepatocyte nuclear factor 1 α (HNF1 α) mRNA and modulates viral replication which negatively regulates HBV enhancer I [35]. Further, miR-548 promoted HBV infection and limits the host response to virus by targeting IFN- λ 1 [36].

Epigenetic modifications of HBV cccDNA are also crucial for viral replication. miR-548ah increases replication of HBV in human hepatoma cells and HBV mouse model [37]. miR-548ah targeted HDAC4, which inhibits deacetylation of histones and its binding to cccDNA thereby promoting viral replication. Recently, Moon et al. (2019) show that overexpression of miR-20a increased the methylation of HBV cccDNA in HepAD38 hepatoma cell line and suppressed HBV replication [38].

Several research groups have identified miRNAs associated with HBV-related carcinogenesis. Fan et al. (2011) observed decrease in miR-122 in HBV-expressing HCC cell line. miR-122 has been shown to target NDRG3 (N-myc downstream-regulated gene 3). Therefore, miR-122 leads to enhanced NDRG3 expression resulting in malignant phenotype [39]. In a different study, lower levels of miR-122 induced the expression of pituitary tumor-transforming gene 1 (PTTG1) binding factor (PBF) which promoted proliferation of HCC cells [40]. Similarly, lower levels of let-7a led to increased cellular growth by targeting signal transducer and activator of transcription 3 (STAT3) leading to HCC [41]. It has been demonstrated that miR-155 upregulated IFN-inducible genes in human hepatoma cell line and repressed HBV disease progression by targeting suppressor of cytokine signaling 1 (SOCS1) which ultimately activated JAK/STAT signaling [42].

Additionally, several miRNAs are also implicated in HBV-associated liver fibrosis/cirrhosis. Lakner and colleagues (2012) reported that miR-19b significantly inhibited TGF- β signaling in activated hepatic satellite cells (HSCs) and consequently protected the cells from fibrosis [43]. Similarly, Venugopal et al. (2010) reported reduction in levels of miR-150 and miR-194 in fibrosis-affected HSCs. These miRNAs target c-myc and rac1 and inhibit HSC activation and ECM production [44].

10.4.1.2 Herpesvirus and miRNAs

Herpesviruses are dsDNA viruses of family *Herpesviridae*. There are eight herpesviruses which are categorized into three subfamilies based on their sequence

similarity: (1) *Alphaherpesvirinae*, (2) *Betaherpesvirinae*, and (3) *Gammaherpesvirinae* [45]. Among all the viral-encoded miRNAs currently known, the majority is identified in herpesviruses (Table 10.2).

Alphaherpesvirus

Herpes simplex virus 1 and 2 (HSV-1 and HSV-2) serotypes of alpha herpesvirus infect humans primarily through oral or genital contact, causing cold sores and genital herpes. The virus migrates from the site of infection and enters neuronal cells, where it establishes a latent infection [46].

HSV-1 and HSV-2 transcribe 18 pre-miRNAs that, respectively, form 27 and 24 different mature miRNAs. hsv-miR-H3 and hsv-miR-H4 target infected cell polypeptide 34.5 (ICP34.5), a lytic neurovirulence factor [47]. hsv-miR-H6 targets infected cell polypeptide 4 (ICP4) protein and promotes maintenance of latent state [48, 49]. Normally, ICP4 downregulates LAT and promotes virus toward lytic infection. Another HSV-1 miRNA, hsv-miR-H2 targets infected cell polypeptide 0 (ICP0) involved in lytic infection and allows the viral entry into replication cycle [50, 51].

Besides targeting viral mRNAs, HSV-1 also encodes hsv-miR-H27 that targets Kelch-like 24 (KLHL24) and prevents transcription of immediate-early and early genes of the virus [52]. This promotes efficient replication and proliferation of HSV-1. HSV-2 miRNAs are homologous to HSV-1 miRNAs. Thus, both HSV-1 and HSV-2 miRNAs play a similar role in viral latency [47, 52].

Similarly, upon HSV-1 infection, cellular miRNAs are also expressed that are considered to be involved in antiviral activities. Cellular miRNA, miR-101, has been shown to target a cellular protein, mitochondrial ATP synthase subunit beta (ATP5B), involved in viral infection. Thus, miR-101/ATP5B might provide cellular protection and inhibit viral replication [53]. HSV-1 also induces miR-146a, which targets complement factor H and activates arachidonic acid cascade implicated in pathological changes associated with neurodegenerative disorders [54].

Betaherpesvirus

Human cytomegalovirus (HCMV) is a DNA virus of β -herpesvirus subfamily. HCMV causes persistent infection in immunocompromised patients and is a major cause of congenital abnormalities. To date, ~26 mature miRNAs encoded by HCMV have been identified along with their potential targets. The genes of these miRNAs are found to be dispersed throughout the viral genome [55, 56].

Both viral and cellular genes are targeted by HCMV-encoded miRNAs. Upon viral infection, hcmv-miR-UL112 targets the MHC-I-related chain B (MICB) and reduces binding of natural killer group 2, member D (NKG2D), hence protecting HCMV from killing by natural killer (NK) cells [57]. Importantly, hcmv-miR-UL112 acts synergistically with hsa-miR-376a and suppresses the expression of MHC class I polypeptide-related sequence B (MICB) and NK cell-mediated killing [58]. Various viral and cellular genes have been shown to be targeted by HCMV-encoded miRNAs, namely, hcmv-miR-UL-112-1, US25-1, US25-2, US25-2-5p, US5-1, US33-5p, and ULD148D, and result in inhibition of viral DNA replication.

Table 10.2 Summary of miRNAs involved in herpesvirus infection

Virus	miRNA	Target gene(s)	Functions
HSV	miR-H2	ICP0	Promotes entry of HSV-1 in replication cycle
	miR-H3 and miR-H4	ICP34. 5	Neurovirulence determinant
	miR-H6	ICP4	Promotes maintenance of latent state
	miR-H27	KLHL24	Efficient replication and proliferation of HSV
HCMV	miR-UL112	MICB	Protects killing of HCMV by NK cells
	miR-UL112-1	IL-32	Alters innate as well as adaptive immune responses
	miR-US25-2-3p	TIMP3	Decreased recognition by NK cells
	miR-US4-1	ERAP1	Suppresses CD8+ T-cell immune responses
	miR-UL112-3p	TLR2	Mediates NF- κ B signaling
	miR-UL112-1, miR-US5-1, and miR-US5-2	SNAP23, VAMP3, RAB11A, and RAB5C	Facilitate formation of VAC for efficient production of virus particles
	miRUS25-1-5p	YWHAE, UBB, NPM1, and HSP90AA1	Suppresses HCMV replication
EBV	miR-BHRF1-3, miR-BART2-5p, miR-BART15	CXCL11, MICB, and NLRP3, respectively	Modulate immune responses
	miR-BART5-5p, miR-BART1, miR-BART16, miR-BART1-3p and miR-BART20-5p	PUMA, BIM, TOMM22, caspase-3, and BAD, respectively	Inhibit apoptosis at early stages of infection
	miR-BART3-5p, miR-BART19-3p	DICE1, WIF1	Promote B-cell transformation and proliferation
	miR-BART7 and miR-BART19-3p	APC	Promotes cellular proliferation
	miR-BART14, miR-BART18-5p, miR-BART19-3p	Nlk	Promotes Wnt signaling
	miR-BART-7-3p	PTEN	Epithelial to mesenchymal transition
KSHV	miR-K12-1	MICB	Promotes growth of KSHV-infected cells
	miR-K12-3 and miR-K12-7	C/EBP β	Stimulate growth of KSHV-infected cells
	miR-K12-5 and miR-K12-9	MYD88 and IRAK1	Signal TLR/IL-1R reduce inflammation

(continued)

Table 10.2 (continued)

Virus	miRNA	Target gene(s)	Functions
	miR-K12-10	TWEAKR	Reduces expression of IL-8 and MCP-1

ICP0, infected cell polypeptide 0, *ICP34.5* infected cell polypeptide 34.5, *ICP4* infected cell polypeptide 4, *KLHL24* Kelch-like 24, *MICB* major histocompatibility complex class I-related chain B, *IL-32* interleukin-32, *TIMP3* tissue inhibitors of metalloprotease 3, *ERAP1* endoplasmic reticulum aminopeptidase1, *TLR2* Toll-like receptor 2, *SNAP23* synaptosomal-associated protein, 23 kDa, *VAMP3* vesicle-associated membrane protein 3, *RAB5C* RAS-related protein 5C, *RAB11A* RAS-related protein 11A, *YWHAE* tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon, *UBB* ubiquitin, *NPM1* nucleoplasmin 1, *HSP90AA1* heat-shock protein 90 alpha family class A member 1, *CXCL11* C-X-C motif chemokine 11, *MICA* MHC class I polypeptide-related sequence B, *NLRP3* NLR family pyrin domain containing 3, *PUMA* p53-upregulated modulator of apoptosis, *BIM* BCL2 interacting mediator of cell death, *C/EBPβ* CCAAT/enhancer-binding protein β, *MYD88* myeloid differentiation primary response 88, *IRAK1* interleukin-1 receptor-associated kinase 1, *TWEAKR* tumor necrosis factor-like weak inducer of the apoptosis receptor, *TOMM22* translocase of outer mitochondrial membrane 22 homolog, *BAD* BCL2-associated death promoter, *DICE1* deleted in cancer 1, *WIF1* WNT inhibitory factor 1, *APC* adenomatous polyposis coli, *Nik* nemo-like kinase, *PTEN* phosphatase and tensin homolog

The viral miRNAs, hcmv-miR-UL112-1 and hcmv-miR-US4-1, regulate immune responses by targeting cytotoxic T lymphocytes (CTL) and NK cells. hcmv-miR-UL112-1 targets 3'UTR of interleukin-32 mRNA and modulates immune responses [59]. hcmv-miR-US25-2-3p reduces the expression of tissue inhibitors of metalloprotease 3 (TIMP3), thereby promoting shedding of MHC-I-related chain A (MICA), and hence decreases recognition by NK cells [60]. hcmv-miR-US4-1 targets aminopeptidase ERAP1 and inhibits CD8⁺ T-cell-mediated immune responses [61].

Similarly, HCMV miRNAs hcmv-miR-UL112-3p, US5-1, UL112-1, and US25-1-5p have been shown to target multiple host inflammatory genes. During HCMV infection, hcmv-miR-UL112-3p has been shown to target Toll-like receptor 2 (TLR2), resulting in inhibition of TLR2-mediated NF-κB signaling [62]. This provides an efficient way of regulating the innate immune response by viral miRNA. HCMV miRs UL112-1, US5-1, and US5-2 target host endocytic machinery. The main target genes include SNAP23 (synaptosomal-associated protein, 23 kDa), VAMP3 (vesicle-associated membrane protein 3), RAB11A (RAS-related protein 11A), and RAB5C (RAS-related protein 5C). These miRNAs coordinately interfere with secretion of pro-inflammatory cytokines, facilitating the production of virion assembly compartment (VAC) resulting in efficient formation of virus particles [63]. hcmv-miRUS25-1-5p targets both viral genes including IE72 and pp65 and host genes including nucleoplasmin 1 (NPM1), ubiquitin (UBB), and tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon (YWHAE) and affects viral replication [64]. Thus, HCMV and cellular miRNAs together play a significant role in promoting latent infection.

Gammaherpesvirus

Like other herpesviruses, Epstein-Barr virus (EBV) also establishes latent infection that persists throughout the life and is usually related to lymphocyte proliferative disorders like Burkitt's lymphoma and Hodgkin's lymphoma. The first viral-encoded miRNAs were identified in EBV [65]. There are three clusters in which viral miRNAs are dispersed: BHRF1-cluster and BART-cluster 1 and 2. EBV transcribes 25 pre-miRNAs that form 44 mature miRNAs. EBV miRNAs target viral as well as cellular mRNAs and help immune evasion and inhibit apoptosis, promoting tumorigenesis, cellular proliferation, and transformation.

ebv-miR-BHRF1-3, ebv-miR-BART2-5p, and ebv-miR-BART15 directly target cellular mRNAs, including CXCL11 (C-X-C motif chemokine 11), MICB, and NLRP3 (NLR family pyrin domain containing 3), respectively, and modulate the immune response in EBV-associated pathologies [66, 67]. The levels of cellular pro-apoptotic genes, including PUMA (p53-upregulated modulator of apoptosis), TOMM22 (translocase of outer mitochondrial membrane 22 homolog), caspase-3, BAD (BCL2-associated death promoter), and BIM (BCL2 interacting mediator of cell death), are modulated by viral miRNAs, ebv-miR-BART5-5p, ebv-miR-BART16, ebv-miR-BART1-3p, ebv-miR-BART20-5p, and ebv-miR-BART1, respectively [68–71]. These viral miRNAs inhibit apoptosis at initial phases of infection.

Viral miRNAs also target tumor-suppressor genes and promote B-cell transformation and proliferation. ebv-miR-BART3-5p inhibits DICE1 (deleted in cancer 1), resulting in enhanced proliferation of cells in vitro [72]. Other tumor-suppressor genes targeted by viral miRNA, ebv-miR-BART19-3p, include WIF1 (WNT inhibitory factor 1), whereas ebv-miR-BART7 and ebv-miR-BART19-3p target APC (adenomatous polyposis coli) and ebv-miR-BART14, ebv-miR-BART18-5p, and ebv-miR-BART19-3p target nemo-like kinase (Nik) key inhibitory genes of Wnt pathway [73]. Cai et al. (2015) showed that ebv-miR-BART-7-3p, highly expressed in nasopharyngeal carcinoma, targeted tumor-suppressor phosphatase and tensin homolog (PTEN) tumor-suppressor gene, regulating PI3K/Akt/GSK-3 β signaling pathway and thereby causing epithelial to mesenchymal transition and metastasis [74]. Altogether, these evidences suggest the role of EBV miRNAs in oncogenesis.

Another member of gammaherpesvirus subfamily, Kaposi's sarcoma-associated herpesvirus (KSHV), encodes 25 mature miRNAs, originating from 13 pre-miRNAs, dispersed disproportionately all over the viral genome. Like EBV, KSHV miRNAs target both viral and cellular mRNAs and modulate immune evasion, avoid apoptosis, and promote tumorigenesis.

KSHV miRNA, kshv-miR-K12-1, targets MICB like HCMV and EBV; kshv-miR-K12-3 and kshv-miR-K12-7 target C/EBP β (CCAAT/enhancer-binding protein β), which transcriptionally represses the expression of IL-6 and IL-10, thus promoting the growth of KSHV-infected cells, and angiogenesis [75]; kshv-miR-K12-5 targets MYD88 (myeloid differentiation primary response 88), whereas kshv-miR-K12-9 targets IRAK1 (interleukin-1 receptor-associated kinase 1). IRAK1 and MYD88 signal TLR/IL-1R and reduce inflammation [76]. kshv-miR-K12-10 represses TWEAKR (tumor necrosis factor-like weak inducer of the apoptosis

receptor) and thus leads to decreased expression of IL-8 and monocyte chemoattractant protein 1 (MCP-1) [77].

Deregulated expression of cellular miRNAs can lead to KSHV-related cancers, regulation of the immune response, and viral replication cycle. A viral protein, K15, upregulates the expression of hsa-miR-21 and hsa-miR-31, thereby promoting cell migration and angiogenesis, suggestive of a significant role of K15 in KSHV-mediated tumor metastasis and cancer [78]. hsa-miR-146a reduces the expression of CXCR4 and is upregulated by viral FLICE inhibitory protein (vFLIP). This promotes early release of viral-infected cellular progenitors into the blood and contributes to Kaposi's sarcoma [79].

10.4.2 RNA Viruses

The miRNAs are also responsible for the establishment of viral infection in RNA viruses. The major effects produced by miRNAs in RNA viruses, for example, hepatitis C virus, human immunodeficiency virus, influenza virus, and dengue virus, have been summarized in Fig. 10.3.

10.4.2.1 Hepatitis C Virus and miRNAs

The hepatitis C virus (HCV) is a member of the *Flaviviridae* family with its genome made up of a single-stranded, positive-sense RNA. Acute and chronic liver infection is caused by HCV, affecting 2–3% of the population worldwide. Most infected patients get hepatocyte infection that eventually leads to cirrhosis of the liver and HCC.

After the virus enters the host cell, its RNA is translated into a viral polyprotein, which then undergoes posttranslational modifications by both viral and host proteases into one core protein, two envelope glycoproteins (E1 and E2), and nonstructural proteins (P7 ion channel, NS2, NS3, NS4A, NS4B, NS5A, and NS5B). The viral genome is flanked by 5' and 3'UTR [80]. Several miRNAs are known to interact directly with the HCV genome (Table 10.3).

For instance, miR-122 is expressed predominantly in hepatocytes and interacts with 5'UTR of viral genome and stimulates virus replication and survival. A stable heterotrimeric complex is formed by the interaction between two molecules of miR-122 at two sites within the 5'UTR. miR-122 also associates with AGO2 protein and thus protects RNA from degradation by host 5' exonuclease Xrn1, thereby stabilizing the viral RNA. It is also demonstrated that interaction of miR-122 with viral RNA at 5'UTR generates 3' overhang which hides the 5'UTR and thus prevents identification by RNA helicase and reduces its degradation [81, 82]. Therefore, miR-122 plays a pivotal role in HCV replication and survival.

miR-199a binding occurs at a site downstream of miR-122 within the 5'UTR. Increased expression of miR-199a inhibits viral replication in two cell lines having HCV-1b or HCV-2a replicons and thus counteracts the action of miR-122 [83]. Similarly, Let-7b binds to a site in 5'UTR and two sites in NS5B coding region. Let-7b suppresses HCV replication as its binding to the viral genome reduces RNA

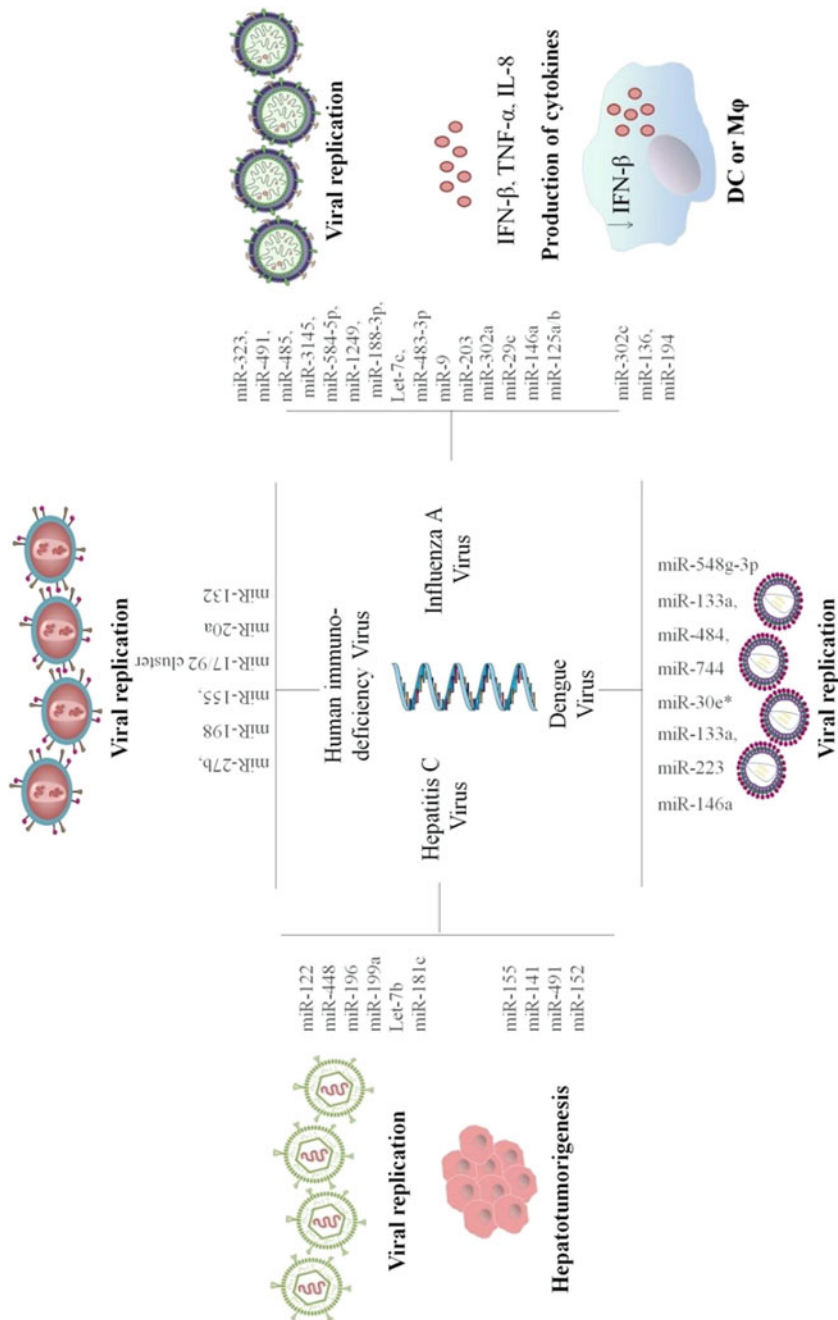


Fig. 10.3 RNA viruses and miRNAs. In hepatitis C virus, dengue virus, influenza virus, and HIV, certain miRNAs promote viral infection by modulating viral replication, promoting cellular proliferation, and altering cytokine production

Table 10.3 Summary of miRNAs involved in HCV-related complications

miRNA	Expression after infection	Target gene(s)	Functions
miRNAs that directly bind HCV genome and modulate viral infection			
miR-122	Up	Unknown	Stabilizes viral RNA and promotes survival
miR-448	Up	IFN- β	Mediates IFN- β -induced antiviral responses against HCV
miR-196	Down	Bach1 mRNA	Promotes upregulation of HMOX1 and enhances antiviral effects against HCV
miR-199a	Up	Unknown	Inhibits HCV replication
Let-7b	Up	Unknown	Negatively regulates HCV replication
miR-181c	Down	HOXA1, STAT3, and STAT5	Upregulates HOX1 expression and its downstream targets STAT3 and STAT5, required for hepatocyte growth
miRNAs implicated in HCV-associated HCC			
miR-155	Up	APC	Activates Wnt/ β -catenin signaling
miR-141	Up	DLC-1	Hepatic tumorigenesis
miR-491	Down	Unknown	Inhibits PI3K/Akt pathway
miR-152	Down	Wnt1	Activates Wnt signaling, subsequently promoting liver tumorigenesis

Up upregulated, *Down* downregulated, *IFN- β* interferon- β , *Bach1* BTB and CNC homology 1, *HOXA1* homeobox A1, *STAT3 and STAT5* signal transducer and activator of transcription 3 and 5, *APC* adenomatous polyposis coli, *DLC-1* deleted in liver cancer

accumulation. However, this suppression does not affect viral translation though its mechanism of action is not clear [84]. miR-196 and miR-448 bind to a site in the NS5A coding region and structural core coding region, respectively. miR-196 binding enhances the antiviral effect by modifying the expression of cellular genes, such as heme oxygenase-1 (HMOX1) and Bach1. It promotes upregulation of HMOX1 and suppression of Bach1 by directly binding at the 3'UTR of Bach1 mRNA [85]. miR-448 causes inhibition of viral replication in Huh cells and mediates IFN- β induced antiviral responses against HCV [86]. miR-181c targets E1 and NS5A regions of the viral genome and is downregulated following HCV infection in hepatocytes. Homeobox A1 (HOXA1) is the direct target of miR-181c. miR-181c downregulation induces HOXA1 expression and its downstream molecules and STAT3 and STAT5 (signal transducer and activator of transcription 3 and 5), which are essential for hepatocyte growth [87]. These cellular miRNAs directly bind the viral genome and exhibit antiviral activity.

Several other miRNAs, including miR-155, miR-141, miR-491, and miR-152, play a vital role in HCV-associated HCC. In patients with chronic HCV infection, a significant increase in miR-155 levels was found in serum and peripheral blood mononuclear cells (PBMCs). HCV core and NS3 and NS5 proteins cause a marked increase in miR-155 levels and TNF α production in human monocytes [88]. Overexpression of miR-155 induces proliferation of hepatocytes and

tumorigenesis by activation of Wnt/ β -catenin signaling in vivo and in vitro and inhibits cellular apoptosis, whereas inhibition of miR-155 induces G₀/G₁ arrest. Adenomatous polyposis coli (APC), which negatively regulates Wnt signaling, is a direct target of miR-155 [89]. miR-141 negatively regulates DLC-1 (deleted in liver cancer) in HCV-infected cells and favors viral replication resulting in increased cell proliferation, hence hepatotumorigenesis [90].

miR-491 and miR-152 function as tumor suppressor, and their reduced expression promotes HCV-associated tumorigenesis. miR-152 levels are regulated by HCV core protein, a known oncoprotein implicated in HCV-related HCC. It has been shown to significantly reduce miR-152 expression leading to activation of Wnt signaling pathway, which ultimately promotes hepatic tumorigenesis [91]. Reduced levels of miR-49 inhibit phosphoinositol-3 (PI3) kinase/Akt pathway, a pro-survival pathway in chronic HCV infection, and promote tumorigenesis [92]. However, increased levels of miR-491 induce apoptosis by targeting Bcl-X_L, an anti-apoptotic Bcl-2 family protein, commonly overexpressed in HCC [93].

10.4.2.2 Human Immunodeficiency Virus 1 and miRNAs

The human immunodeficiency virus (HIV) is a virus of the *Retroviridae* family. Its genome comprises of single-stranded RNA. HIV disease has different stages of infection: acute, asymptomatic chronic eventually leading to symptomatic HIV infection or acquired immunodeficiency syndrome (AIDS). Infection with HIV causes a progressive decline in CD4⁺ T lymphocytes, impairing cell-mediated immunity. Host miRNA expression profiles have been shown to be altered in response to HIV-1 infection (Table 10.4).

Several miRNAs expressed by CD4⁺ T lymphocytes target viral accessory genes, viral protein r (vpr), viral infectivity factor (vif), viral protein U (vpu), and negative regulatory factor (nef), involved in viral infection and replication. The cellular miRNAs, mir-149, mir-324-5p, and mir-378, target vpr, vif, and vpu, respectively, while mir-29a and mir-29b target viral nef gene [94]. Another set of miRNAs such as miR-28, miR-125b, miR-150, miR-223, and miR-382 interact directly with 3'UTR and inhibit translation of crucial proteins, Tat and Rev, involved in transcription and translocation of viral RNA. These miRNAs are upregulated in resting CD4⁺ T cells which upon activation cause downregulation of these particular miRNAs. Thus, expression of cellular miRNAs determines the host susceptibility to HIV-1 infection [95].

In addition, cellular miRNAs regulate host factors like cyclin T1, MeCP-2, SIRT1, and PNUTS and thus modulate viral replication. Cyclin T1 binds viral transactivator protein Tat and is required for activation of RNA polymerase II transcription of integrated provirus. miR-27b directly targets 3'UTR of cyclin T1 mRNA and inhibits its expression in resting CD4⁺ T cells, thereby impairing viral replication. Activation of CD4⁺ T cells results in downregulation of miR-27b, restoring cyclin T1 levels and thus increasing the vulnerability of CD4⁺ T cells to infection [96]. Overexpression of miR-198 also represses cyclin T1 levels in promonocytic cell line MM6 and thus restricts HIV-1 replication [97].

Table 10.4 Summary of miRNAs that target host cellular factors in HIV-1 infection

miRNA	Expression after infection	Target gene(s)	Functions
miR-27b, miR-198	Up	Cyclin T1	Prevents activation of CD4+ T cells and thus restricts viral replication
miR-155	Up	LEDGF/p75, ADAM10, nucleoporin, TNPO3	Promotes HBV replication
miR-15a, miR-15b, miR-16, miR-20a, miR-106b, and miR-93	Down	Pur- α	Decreases susceptibility to infection
miR-17/92 cluster and miR-20a	Down	PCAF	Increases viral replication
miR-132	Up	MeCP2	Increases HIV-1 replication
miR-34a	Up	PNUTS	Inhibits HIV-1 transcription
miR-217 and miR-34a	Up	SIRT1	Promotes HIV-1 Tat-mediated transactivation
miR-182	Up	NAMPT	Promotes HIV-1 Tat-mediated transactivation

Up upregulated, *Down* downregulated, *HDFs* HIV-1 dependency factors, *LEDGF* lens epithelium-derived growth factor, *ADAM10* a protein of disintegrin and metalloprotease family, *TNPO3* transportin3, *Pur- α* purine-rich element binding protein α , *PCAF* p300-CREB binding protein-associated factor, *MeCP2* methyl-CpG binding protein 2, *PNUTS* phosphatase 1 nuclear-targeting subunit, *SIRT1* sirtuin-1, *NAMPT* nicotinamide phosphoribosyltransferase

miR-155 has been shown to employ anti-HIV activity. Upon TLR3 and TLR4 stimulation in macrophages, miR-155 levels are upregulated. miR-155 targets 3'UTR of mRNAs of HIV-1 dependency factors (HDFs), lens epithelium-derived growth factor (LEDGF)/p75, ADAM10 (a protein of disintegrin and metalloprotease family), nucleoporin, and TNPO3 (transportin3), engaged in nuclear trafficking and import of pre-integration complexes (PICs). This results in a reduction of mRNA and protein levels of HDFs in monocyte-derived macrophages (MDMs), thereby suggesting a possible role for its novel anti-HIV-1 effect [98]. Another HIV-1 dependency factor, Pur- α (purine-rich element binding protein α), binds HIV-1 TAR element and viral Tat protein and facilitates viral transcription. Cellular miRNAs like miR-15a, miR-15b, miR-16, miR-20a, miR-106b, and miR-93 bind 3'UTR of Pur- α mRNA and modulate viral susceptibility in monocytes [99].

HIV-1 infection significantly reduces the expression of miR-17/92 cluster and miR-20a in Jurkat cell line by positively regulating the expression of p300-CREB binding protein-associated factor (PCAF), a cofactor for Tat acetylation, thereby resulting in enhanced viral replication [100]. miR-132 is highly upregulated during CD4⁺ T-cell activation and enhances viral replication in Jurkat CD4⁺ T-cell line. Methyl-CpG binding protein 2 (MeCP2) is reported as a target of miR-132. miR-132

overexpression promotes downregulation of MeCP2 and increases HIV-1 replication [101].

miR-34a targets phosphatase 1 nuclear-targeting subunit (PNUTS) which results in inhibition of HIV-1 transcription by disrupting the assembly of HIV-1 transcription machinery [102]. Tat-induced upregulation of miR-34a and miR-217 has been observed in TZM-bl cells and MAGI cells, respectively. These miRNAs target 3'UTR of sirtuin-1 (SIRT1) mRNA and downregulate its expression. This promotes acetylation and activation of NF- κ B, which further enhances HIV-1 transactivation [103, 104]. Further, miR-182 downregulates the expression of nicotinamide phosphoribosyltransferase (NAMPT), involved in Tat-induced inhibition of SIRT1 and HIV-1 transactivation [105].

Upon HIV-1 infection in human Sup-T1 cells, the expression of miR-186, miR-210, and miR-222 is upregulated which further inhibits viral replication by downregulating host target genes such as Dicer1, HIV-1 Rev-binding protein, and enhancer binding protein 2. Thus, human miR-186, miR-210, and miR-222 directly regulate key genes involved in HIV-1 replication and miRNA biogenesis [106]. Human T lymphocytes upon HIV-1 infection have been shown to highly express miR-29a which interacts with 3'UTR. This interaction enhances viral interaction with RISC proteins and P bodies and represses viral replication. However, depletion of P-bodies resulted in the enhanced translation of viral proteins and replication [107]. This provides an understanding of how host miRNA levels regulate HIV-1 infection.

10.4.2.3 Influenza Virus and miRNAs

Influenza virus is an RNA virus of family *Orthomyxoviridae*. Of known genera of this family, genus A, B, C, and D, influenza A virus (IAV) is responsible for respiratory illness in birds and mammals. The eight RNA segments transcribe ten viral proteins including acidic polymerase (PA), basic polymerase 1 (PB1), basic polymerase 2 (PB2), hemagglutinin (HA), matrix capsid protein 1 (M1), matrix capsid protein 2 (M2), neuraminidase (NA), nonstructural protein (NS1), nuclear export protein (NEP), and nucleoprotein (NP). Of these, HA and NA transmembrane glycoproteins act as major antigens. On the basis of these surface proteins, there are different HA (18) and NA subtypes (11) in influenza A virus.

Host cellular miRNAs are shown to control IAV life cycle by directly targeting either viral genome, innate immunity, or genes involved in viral replication (Table 10.5). Many cellular miRNAs like miR-323, miR-491, miR-654 [108], miR-485 [109], and miR-3145 [110] bind directly to PB1 gene and inhibit IAV replication. Similarly, miR-584-5p and miR-1249 have been shown to bind directly to PB2 gene and impede viral replication [111]. In a recent study, miR-188-3p downregulated PB2 expression and effectively inhibited IAV replication in A549 cells [112]. Further, overexpression of Let-7c was observed in IAV-infected human lung epithelial (A549) cells and was found to reduce M1 (+) cRNA and suppress viral replication [113].

Several cellular miRNAs modulate viral replication indirectly by targeting the genes involved in immune pathways. For instance, miR-302a modulates

Table 10.5 Summary of miRNAs modulating IAV infection

miRNA	Target gene(s)	Functions
miRNAs bind directly at IAV genome		
miR-323, miR-491, miR-654, miR-485, and miR-3145	PB1	Inhibit IAV replication
miR-584-5p, miR-1249, and miR-188-3p	PB2	Inhibit viral replication
Let-7c	M1 (+) cRNA	Suppresses viral replication
miR-323, miR-491, miR-654, miR-485, and miR-3145	PB1	Inhibit IAV replication
miRNAs indirectly target genes involved in immune signaling pathways		
miR-302a	IRF-5	Stimulates the secretion of IFN- β , TNF α , IL-8, IL-6, CCL2, and CCL5
miR-29c	NF- κ B	Reduces the synthesis of several antiviral and proinflammatory cytokines
miR-146a	TRAF6	Negatively regulates TLR signaling pathway
miR-7, miR-132, miR-146a, miR-187, miR-200c, miR-1275	IRAK1 and MAPK3	Downregulates the antiviral proteins, IRAK1 and MAPK3
miR-125a/b	A20 deubiquitinase	Increases NF- κ B activity and synthesis of pro-inflammatory cytokines
miR-302c	NIK	Prevents the import of NF- κ B from cytoplasm to nucleus and reduces IFN- β production
miR-136 and miR-194	RIG-1	Inhibit the production of IFN- β
miRNAs target genes required for viral replication		
miR-483-3p	RNF5 and CD81	Decreases viral replication
miR-33a	COPI subunit ARCNI	Impairs viral replication at the stage of virus internalization
miR-9	MCP1P1	Promotes IAV replication
miR-203	DR1	Inhibits IAV replication

PB1 basic polymerase1, *PB2* basic polymerase 2, *M1* matrix capsid protein 1, *IRF-5* interferon regulatory factor-5, *NF- κ B* nuclear factor kappa-light-chain-enhancer of activated B cells, *TRAF6* tumor necrosis factor receptor (TNFR)-associated factor 6, *IRAK1* interleukin-1 receptor-associated kinase 1, *MAPK3* mitogen-activated protein kinase 3, *NIK* NF- κ B-inducing kinase, *RIG1* retinoic acid-inducible gene I, *RNF-5* RING-finger protein 5, *COPI* coat protein 1, *ARCNI* archaic 1, *MCP1P1* monocyte chemoattractant protein 1-induced protein, *DR1* downregulator of transcription 1

IAV-induced cytokine storm. It binds directly at 3'UTR of interferon regulatory factor-5 (IRF-5) mRNA and suppresses IRF-5-stimulated cytokine release. However, IAV infection downregulated the expression of miR-302a and upregulated IRF-5 expression. This stimulated the secretion of IFN- β , TNF α , IL-8, IL-6, CCL2, and CCL5 and promoted IAV replication [114]. In contrast, IAV infection induced the expression of miR-29c, which decreased NF- κ B activity and reduced the secretion of many antiviral and pro-inflammatory cytokines in IAV-infected A549 cells

[115]. In IAV H3N2-infected human nasal epithelial cells (hNECs), the induction of miR-146a has been described and is shown to negatively regulate TLR signaling pathway by targeting TRAF6 [116]. In the same line, miR-144 has been shown to reduce the antiviral effect by attenuating TRAF6-IRF7 pathway [117]. In addition, miR-7, miR-132, miR-146a, miR-187, miR-200c, and miR-1275 have been identified to accumulate in A549 cells in response to IAV infection which ultimately downregulates the antiviral proteins like IRAK1 (interleukin-1 receptor-associated kinase 1) and MAPK3 (mitogen-activated protein kinase 3) [118].

The expression of miR-125a/b is elevated during IAV H1N1 infection. miR-125a/b directly targets A20 deubiquitinase, which inhibits NF- κ B activity. Therefore, infection with IAV increased NF- κ B activity and pro-inflammatory cytokine secretion [119] which further lessen antiviral responses. Gui et al. (2015) observed downregulation of miR-302c expression in IAV H3N2-infected cells. miR-302c has been shown to target NF- κ B-inducing kinase (NIK) and reduced IFN- β production. miR-302c also inhibited the translocation of NF- κ B from cytoplasm to nucleus [120].

The pattern recognition receptor retinoic acid-inducible gene I (RIG-I) is important for type-I interferon response after recognizing the cells that have been infected with influenza virus. miR-136 is an immune agonist of RIG-I, resulting in accumulation of IL-6 and IFN- β in IAV H5N1-infected A549 cells. Therefore, miR-136 serves as an activator of the immune system and inhibitor of viral replication in vitro [121]. Furthermore, miR-483-3p was highly expressed in bronchoalveolar lavage fluid (BALF) exosomes from infected mice and enhanced the production of IFN- β and pro-inflammatory cytokines in H1N1-, H7N9-, or H5N1-infected mouse lung epithelial cells. miR-483-3p was found to target RING-finger protein 5 (RNF5) and CD81, the negative mediators of RIG-I pathway resulting in decreased viral replication [122]. Contrarily, miR-194 facilitated viral replication by inhibiting the production of IFN- β via targeting RIG-I pathway in H1N1-infected A549 cells [123].

Several miRNAs have been shown to exert inhibitory effect on viral replication by targeting the genes required for viral replication. For example, miR-33a targets 3'UTR of COPI subunit ARCNI (archain 1) and impairs viral replication at the phase of virus internalization. It also decreases viral ribonucleoprotein activity via ARCNI-independent manner. Therefore, miR-33a both acts as inhibitor of viral replication and interferes with viral internalization [124]. In H1N1- or H3N2-infected A549 cells, miR-9 has been shown to promote IAV replication by repressing monocyte chemoattractant protein 1-induced protein (MCP1P1) that degrades viral RNA and also inhibits viral replication by reducing the production of viral NP and M proteins [125]. In H5N1-infected A549 cells, miR-24 regulates furin-mediated proteolytic activation of HA0 glycoproteins and production of infectious virions [126]. Recently, Zhang et al. (2018) has demonstrated that upregulation of miR-203 in H5N1-infected A549 cells inhibited IAV replication by targeting downregulator of transcription 1 (DR1) [127].

10.4.2.4 Arboviruses and miRNA

Arboviruses are arthropod-borne viruses which transmit infections in vertebrate hosts through the bite of hematophagous arthropods including mosquitoes, sandflies, and ticks. Arboviruses have RNA genome and include *Flaviviridae*, *Togaviridae*, *Bunyaviridae*, *Rhabdoviridae*, and *Reoviridae* families. Of these, flaviviruses belong to the family *Flaviviridae* which includes most prominent human pathogens like chikungunya virus, dengue virus, West Nile virus, Zika virus, and various others. Here, we will specifically focus on the modulation of miRNA expression in dengue virus (DENV) infection (Table 10.6).

The first evidence for the role of miRNAs in DENV infection was provided by Wen and colleagues (2015). They observed that miR-548 g-3p targeted stem loop A (SLA) promoter present in 5'UTR and suppressed DENV translation, replication, and multiplication [128]. Three miRNAs, miR-133a, miR-484, and miR-744, were found to target viral 3'UTR and inhibit viral replication [129, 130].

Similarly, another set of miRNAs regulate DENV replication indirectly by regulating the host immune system or by targeting the host cellular factors required in the viral life cycle. For instance, Escalera-Cueto et al. (2015) demonstrated that *Let-7c* is highly upregulated in hepatic Huh-7 cells during DENV infection. *Let-7c* targets a transcription factor *Bach1* (BTB and CNC homology 1), repressor of anti-inflammatory protein heme oxygenase-1 [131]. This protects the infected cells from excessive production of oxidative stress and inflammation. Similarly, the highly expressed miRNA, miR-30e*, was shown to modulate the immune response in human monocyte U937 and HeLa cell line by upregulating the expression of IFN- β and downstream genes such as *OAS1* (2'-5'-oligoadenylate synthetase 1), *MxA* (*Myxovirus* resistance gene A) and *IFITM1* (interferon-induced transmembrane protein 1). At the same time, miR-30e* also activated NF- κ B by targeting

Table 10.6 miRNAs associated with dengue virus infection

miRNA	Target gene(s)	Functions
miR-548 g-3p	SLA promoter	Suppresses DENV translation, replication, and multiplication
miR-133a, miR-484, and miR-744	Viral 3'UTR	Inhibit viral replication
<i>Let-7c</i>	<i>Bach1</i>	Protects the infected cells from excessive production of oxidative stress and inflammation
miR-30e*	<i>IκBα</i>	Upregulates the expression of IFN- β and inhibits DENV replication
miR-34a, miR-34c, miR-449a, and miR-449b	Indirectly by Wnt pathway	Positively modulate immune response
miR-133a and miR-223	<i>PTB</i> and <i>STMN1</i>	Suppresses DENV replication
miR-146a	<i>TRAF6</i>	Facilitates DENV replication

SLA stem loop A, *Bach1* BTB and CNC homology 1, *I κ B α* NF-kappa-B inhibitor alpha, *PTB* polypyrimidine tract binding, *STMN1* stathmin 1, *TRAF6* tumor necrosis factor receptor-associated factor 6

I κ B α (NF-kappa-B inhibitor alpha) 3'UTR; this restored the production of IFN- β -mediated antiviral immune response and inhibited DENV replication [132]. Like miR-30e*, miR-34 family including miR-34a, miR-34c, miR-449a, and miR-449b was shown to positively modulate immune response upon DENV infection by increasing the expression of type I IFN and stimulated genes by suppressing Wnt signaling [133].

Other studies provided evidence for involvement of host machinery in modulating DENV multiplication. Castillo et al. (2016) hypothesized that overexpression of miR-133a led to suppression of DENV replication in Vero cells by directly targeting the polypyrimidine tract binding (PTB) protein involved in DENV replication [134]. Like this, miR-223 also inhibited DENV replication by inhibiting stathmin 1 (STMN1), a microtubule-destabilizing protein in human endothelial-like EAhy926 cells [135]. In addition, various other miRNAs have also been reported to facilitate viral replication. Overexpression of miR-146a facilitates DENV replication in monocytic THP-1 cells. miR-146a targets tumor necrosis factor receptor-associated factor 6 (TRAF6), which dampens the secretion of host IFN- β and reduces the antiviral immune responses [136]. HepG2 cells were treated with an anti-miR-21 (AMO-21) before DENV infection, and reduction in DENV production was observed in HepG2 cells, suggesting that miR-21 is involved in DENV replication [137].

10.5 Future Implications: miRNAs for Disease Diagnosis and Therapeutics

miRNAs are expressed ubiquitously in body fluids including peripheral blood, saliva, urine, cerebrospinal fluid (CSF), and other biological samples. The circulating miRNAs remain stable after repeated cycles of freeze-thawing and long-term storage of biological samples [138, 139]. The interaction of miRNAs with target mRNA during host-pathogen interactions provides a tool for deciphering the role of key genes involved in the activation of the immune system and determines the susceptibility of infection. Increasing evidence suggests the relevance of miRNAs as biomarkers for diagnosis of various diseases including cancer and neurodegenerative, cardiovascular, and immune disorders. However, very few studies have elucidated the role of miRNAs as a biomarker for infectious diseases. As in the case of HCV infection, the viral-host interaction can be regulated by targeting cellular miRNA-122. The inhibitor of miR-122, miravirsin, is already in clinical trials and used as a therapeutic strategy against hepatitis C infection [140].

In conclusion, targeting pathogen-encoded miRNAs can be employed as a therapeutic strategy against infectious diseases. This will interfere with the key biological processes, including multiplication or replication of the pathogen. However, translating the miRNAs as biomarkers as well as targets for the treatment of infectious diseases still remains a challenge.

References

1. St Laurent G, Wahlestedt C, Kapranov P (2015) The landscape of long noncoding RNA classification. *Trends Genet* 31(5):239–251
2. Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75(5):843–854
3. Mazière P, Enright AJ (2007) Prediction of microRNA targets. *Drug Discov Today* 12(11–12):452–458
4. Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 19(1):92–105
5. Cullen BR (2011) Herpesvirus microRNAs: phenotypes and functions. *Curr Opin Virol* 1(3):211–215
6. Taganov KD, Boldin MP, Chang KJ, Baltimore D (2006) NF-kappa B-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc Natl Acad Sci U S A* 103(33):12481–12486
7. Ma F, Xu S, Liu X et al (2011) The microRNA miR-29 controls innate and adaptive immune responses to intracellular bacterial infection by targeting interferon- γ . *Nat Immunol* 12(9):861–869
8. Lee Y, Kim M, Han J et al (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* 23(20):4051–4060
9. Lee Y, Jeon K, Lee JT, Kim S, Kim VN (2002) MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J* 21(17):4663–4670
10. Lee Y, Ahn C, Han J et al (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425(6956):415–419
11. Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the microprocessor complex. *Nature* 432(7014):231–235
12. Zeng Y, Yi R, Cullen BR (2005) Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO J* 24(1):138–148
13. Daniels SM, Melendez-Peña CE, Scarborough RJ et al (2009) Characterization of the TRBP domain required for dicer interaction and function in RNA interference. *BMC Mol Biol* 10:38
14. Chendrimada TP, Gregory RI, Kumaraswamy E et al (2005) TRBP recruits the dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* 436(7051):740–744
15. Fang Z, Rajewsky N (2011) The impact of miRNA target sites in coding sequences and in 3'UTRs. *PLoS One* 6(3):e18067
16. Lytle JR, Yario TA, Steitz JA (2007) Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *Proc Natl Acad Sci U S A* 104(23):9667–9672
17. Place RF, Li LC, Pookot D, Noonan EJ, Dahiya R (2008) MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proc Natl Acad Sci U S A* 105(5):1608–1613
18. Lingel A, Simon B, Izaurralde E, Sattler M (2004) Nucleic acid 3'-end recognition by the Argonaute2 PAZ domain. *Nat Struct Mol Biol* 11(6):576–577
19. Yan KS, Yan S, Farooq A, Han A, Zeng L, Zhou MM (2004) Structure and conserved RNA binding of the PAZ domain. *Nature* 427(6971):265
20. Parker JS, Roe SM, Barford D (2004) Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. *EMBO J* 23(24):4727–4737
21. Song JJ, Smith SK, Hannon GJ, Joshua-Tor L (2004) Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* 305(5689):1434–1437
22. Liu J, Carmell MA, Rivas FV et al (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science* 305(5689):1437–1441
23. Meister G, Landthaler M, Patkaniowska A, Dorsett Y, Teng G, Tuschl T (2004) Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell* 15(2):185–197

24. Mathonnet G, Fabian MR, Svitkin YV et al (2007) MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. *Science* 317(5845):1764–1767
25. Chendrimada TP, Finn KJ, Ji X et al (2007) MicroRNA silencing through RISC recruitment of eIF6. *Nature* 447(7146):823–828
26. Teixeira D, Sheth U, Valencia-Sanchez MA, Brengues M, Parker R (2005) Processing bodies require RNA for assembly and contain nontranslating mRNAs. *RNA* 11(4):371–382
27. Eulalio A, Huntzinger E, Izaurralde E (2008) GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. *Nat Struct Mol Biol* 15(4):346–353
28. Zhang GL, Li YX, Zheng SQ, Liu M, Li X, Tang H (2010) Suppression of hepatitis B virus replication by microRNA-199a-3p and microRNA-210. *Antivir Res* 88(2):169–175
29. Potenza N, Papa U, Mosca N, Zerbini F, Nobile V, Russo A (2011) Human microRNA hsa-miR-125a-5p interferes with expression of hepatitis B virus surface antigen. *Nucleic Acids Res* 39(12):5157–5163
30. Qiu L, Fan H, Jin W et al (2010) miR-122-induced down-regulation of HO-1 negatively affects miR-122-mediated suppression of HBV. *BiochemBiophys Res Commun* 398(4):771–777
31. Protzer U, Seyfried S, Quasdorff M et al (2007) Antiviral activity and hepatoprotection by heme oxygenase-1 in hepatitis B virus infection. *Gastroenterology* 133(4):1156–1165
32. Wang S, Qiu L, Yan X et al (2012) Loss of microRNA 122 expression in patients with hepatitis B enhances hepatitis B virus replication through cyclin G(1)-modulated P53 activity. *Hepatology* 55(3):730–741
33. Jin J, Tang S, Xia L et al (2013) MicroRNA-501 promotes HBV replication by targeting HBXIP. *BiochemBiophys Res Commun* 430(4):1228–1233
34. Guo H, Liu H, Mitchelson K et al (2011) MicroRNAs-372/373 promote the expression of hepatitis B virus through the targeting of nuclear factor I/B. *Hepatology* 54(3):808–819
35. Dai X, Zhang W, Zhang H et al (2014) Modulation of HBV replication by microRNA-15b through targeting hepatocyte nuclear factor 1 α . *Nucleic Acids Res* 42(10):6578–6590
36. Li Y, Xie J, Xu X et al (2013) MicroRNA-548 down-regulates host antiviral response via direct targeting of IFN- λ 1. *Protein Cell* 4(2):130–141
37. Xing T, Zhu J, Xian J et al (2019) miRNA-548ah promotes the replication and expression of hepatitis B virus by targeting histone deacetylase 4. *Life Sci* 219:199–208
38. Moon IY, Choi JH, Chung JW, Jang ES, Jeong SH, Kim JW (2019) MicroRNA-20 induces methylation of hepatitis B virus covalently closed circular DNA in human hepatoma cells. *Mol Med Rep* 20(3):2285–2293
39. Fan CG, Wang CM, Tian C et al (2011) miR-122 inhibits viral replication and cell proliferation in hepatitis B virus-related hepatocellular carcinoma and targets NDRG3. *Oncol Rep* 26(5):1281–1286
40. Li C, Wang Y, Wang S et al (2013) Hepatitis B virus mRNA-mediated miR-122 inhibition upregulates PTTG1-binding protein, which promotes hepatocellular carcinoma tumor growth and cell invasion. *J Virol* 87(4):2193–2205
41. Wang Y, Lu Y, Toh ST et al (2010) Lethal-7 is down-regulated by the hepatitis B virus x protein and targets signal transducer and activator of transcription 3. *J Hepatol* 53(1):57–66
42. Su C, Hou Z, Zhang C, Tian Z, Zhang J (2011) Ectopic expression of microRNA-155 enhances innate antiviral immunity against HBV infection in human hepatoma cells. *Virol J* 8:354
43. Lakner AM, Steuerwald NM, Walling TL et al (2012) Inhibitory effects of microRNA 19b in hepatic stellate cell-mediated fibrogenesis. *Hepatology* 56(1):300–310
44. Venugopal SK, Jiang J, Kim TH et al (2010) Liver fibrosis causes downregulation of miRNA-150 and miRNA-194 in hepatic stellate cells, and their overexpression causes decreased stellate cell activation. *Am J PhysiolGastrointest Liver Physiol* 298(1):G101–G106
45. Cullen BR (2011) Herpesvirus microRNAs: phenotypes and functions. *Curr Opin Virol* 1(3):211–215

46. Du T, Han Z, Zhou G, Roizman B (2015) Patterns of accumulation of miRNAs encoded by herpes simplex virus during productive infection, latency, and on reactivation. *Proc Natl Acad Sci U S A* 112(1):E49–E55
47. Tang S, Patel A, Krause PR (2009) Novel less-abundant viral microRNAs encoded by herpes simplex virus 2 latency-associated transcript and their roles in regulating ICP34. 5 and ICP0 mRNAs. *J Virol* 83(3):1433–1442
48. Tang S, Bertke AS, Patel A, Margolis TP, Krause PR (2011) Herpes simplex virus 2 microRNA miR-H6 is a novel latency-associated transcript-associated microRNA, but reduction of its expression does not influence the establishment of viral latency or the recurrence phenotype. *J Virol* 85(9):4501–4509
49. Duan F, Liao J, Huang Q, Nie Y, Wu K (2012) HSV-1 miR-H6 inhibits HSV-1 replication and IL-6 expression in human corneal epithelial cells in vitro. *Clin Dev Immunol* 2012:192791
50. Umbach JL, Kramer MF, Jurak I, Karnowski HW, Coen DM, Cullen BR (2008) MicroRNAs expressed by herpes simplex virus 1 during latent infection regulate viral mRNAs. *Nature* 454(7205):780–783
51. Everett RD (2000) ICP0, a regulator of herpes simplex virus during lytic and latent infection. *BioEssays* 22(8):761–770
52. Umbach JL, Nagel MA, Cohrs RJ, Gilden DH, Cullen BR (2009) Analysis of human alphaherpesvirus microRNA expression in latently infected human trigeminal ganglia. *J Virol* 83:10677–10683
53. Zheng SQ, Li YX, Zhang Y, Li X, Tang H (2011) MiR-101 regulates HSV-1 replication by targeting ATP5B. *Antivir Res* 89(3):219–226
54. Hill JM, Zhao Y, Clement C, Neumann DM, Lukiw WJ (2009) HSV-1 infection of human brain cells induces miRNA-146a and Alzheimer-type inflammatory signaling. *Neuroreport* 20(16):1500–1505
55. Grey F, Antoniewicz A, Allen E et al (2005) Identification and characterization of human cytomegalovirus-encoded microRNAs. *J Virol* 79(18):12095–12099
56. Stark TJ, Arnold JD, Spector DH, Yeo GW (2012) High-resolution profiling and analysis of viral and host small RNAs during human cytomegalovirus infection. *J Virol* 86(1):226–235
57. Stern-Ginossar N, Elefant N, Zimmermann A et al (2007) Host immune system gene targeting by a viral miRNA. *Science* 317(5836):376–381
58. Nachmani D, Lankry D, Wolf DG, Mandelboim O (2010) The human cytomegalovirus microRNA miR-UL112 acts synergistically with a cellular microRNA to escape immune elimination. *Nat Immunol* 11:806–813
59. Huang Y, Qi Y, Ma Y, He R, Ji Y, Sun Z, Ruan Q (2013) The expression of interleukin-32 is activated by human cytomegalovirus infection and down regulated by hcmv-miR-UL112-1. *Virology* 451:10–15
60. Esteso G, Luzón E, Sarmiento E et al (2014) Altered microRNA expression after infection with human cytomegalovirus leads to TIMP3 downregulation and increased shedding of metalloprotease substrates, including MICA. *J Immunol* 193(3):1344–1352
61. Kim S, Lee S, Shin J et al (2011) Human cytomegalovirus microRNA miR-US4-1 inhibits CD8(+) T cell responses by targeting the aminopeptidase ERAP1. *Nat Immunol* 12(10):984–991
62. Landais I, Pelton C, Streblov D, DeFilippis V, McWeeney S, Nelson JA (2015) Human cytomegalovirus miR-UL112-3p targets TLR2 and modulates the TLR2/IRAK1/NFκB signaling pathway. *PLoS Pathog* 11(5):e1004881
63. Hook LM, Grey F, Grabski R et al (2014) Cytomegalovirus miRNAs target secretory pathway genes to facilitate formation of the virion assembly compartment and reduce cytokine secretion. *Cell Host Microbe* 15(3):363–373
64. Jiang S, Qi Y, He R et al (2015) Human cytomegalovirus microRNA miR-US25-1-5p inhibits viral replication by targeting multiple cellular genes during infection. *Gene* 570(1):108–114
65. Pfeffer S, Zavolan M, Grässer FA et al (2004) Identification of virus-encoded microRNAs. *Science* 304(5671):734–736

66. Xia T, O'Hara A, Araujo I et al (2008) EBV microRNAs in primary lymphomas and targeting of CXCL-11 by ebv-mir-BHRF1-3. *Cancer Res* 68(5):1436–1442
67. Haneklaus M, Gerlic M, Kurowska-Stolarska M (2012) Cutting edge: miR-223 and EBV miR-BART15 regulate the NLRP3 inflammasome and IL-1 β production. *J Immunol* 189(8):3795–3799
68. Choy EY, Siu KL, Kok KH et al (2008) An Epstein-Barr virus-encoded microRNA targets PUMA to promote host cell survival. *J Exp Med* 205(11):2551–2560
69. Marquitz AR, Mathur A, Nam CS, Raab-Traub N (2011) The Epstein-Barr virus BART microRNAs target the pro-apoptotic protein Bim. *Virology* 412(2):392–400
70. Vereide DT, Seto E, Chiu YF et al (2014) Epstein-Barr virus maintains lymphomas via its miRNAs. *Oncogene* 33(10):1258–1264
71. Kim H, Choi H, Lee SK (2015) Epstein-Barr virus miR-BART20-5p regulates cell proliferation and apoptosis by targeting BAD. *Cancer Lett* 356(2 PtB):733–742
72. Lei T, Yuen KS, Xu R et al (2013) Targeting of DICE1 tumor suppressor by Epstein-Barr virus-encoded miR-BART3* microRNA in nasopharyngeal carcinoma. *Int J Cancer* 133(1):79–87
73. Wong AM, Kong KL, Tsang JW, Kwong DL, Guan XY (2012) Profiling of Epstein-Barr virus-encoded microRNAs in nasopharyngeal carcinoma reveals potential biomarkers and oncomirs. *Cancer* 118(3):698–710
74. Cai L, Ye Y, Jiang Q et al (2015) Epstein-Barr virus-encoded microRNA BART1 induces tumour metastasis by regulating PTEN-dependent pathways in nasopharyngeal carcinoma. *Nat Commun* 6:7353
75. Qin Z, Kearney P, Plaisance K, Parsons CH (2010) Pivotal advance: Kaposi's sarcoma-associated herpesvirus (KSHV)-encoded microRNA specifically induce IL-6 and IL-10 secretion by macrophages and monocytes. *J Leukoc Biol* 87:25–34
76. Abend JR, Ramalingam D, Kieffer-Kwon P, Uldrick TS, Yarchoan R, Ziegelbauer JM (2012) Kaposi's sarcoma-associated herpesvirus microRNAs target IRAK1 and MYD88, two components of the toll-like receptor/interleukin-1R signaling cascade, to reduce inflammatory-cytokine expression. *J Virol* 86:11663–11674
77. Abend JR, Uldrick T, Ziegelbauer JM (2010) Regulation of tumor necrosis factor like weak inducer of apoptosis receptor protein (TWEAKR) expression by Kaposi's sarcoma-associated herpesvirus microRNA prevents TWEAK induced apoptosis and inflammatory cytokine expression. *J Virol* 84:12139–12151
78. Tsai YH, Wu MF, Wu YH et al (2009) The M type K15 protein of Kaposi's sarcoma-associated herpesvirus regulates microRNA expression via its SH2-binding motif to induce cell migration and invasion. *J Virol* 83(2):622–632
79. Punj V, Matta H, Schamus S, Tamewitz A, Anyang B, Chaudhary PM (2010) Kaposi's sarcoma-associated herpesvirus-encoded viral FLICE inhibitory protein (vFLIP) K13 suppresses CXCR4 expression by upregulating miR-146a. *Oncogene* 29(12):1835–1844
80. Suzuki T, Ishii K, Aizaki H, Wakita T (2007) Hepatitis C viral life cycle. *Adv Drug Deliv Rev* 59(12):1200–1212
81. Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P (2005) Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science* 309(5740):1577–1581
82. Mortimer SA, Doudna JA (2013) Unconventional miR-122 binding stabilizes the HCV genome by forming a trimolecular RNA structure. *Nucleic Acids Res* 41(7):4230–4240
83. Murakami Y, Aly HH, Tajima A, Inoue I, Shimotohno K (2009) Regulation of the hepatitis C virus genome replication by miR-199a. *J Hepatol* 50(3):453–460
84. Cheng JC, Yeh YJ, Tseng CP et al (2012) Let-7b is a novel regulator of hepatitis C virus replication. *Cell Mol Life Sci* 69(15):2621–2633
85. Hou W, Tian Q, Zheng J, Bonkovsky HL (2010) MicroRNA-196 represses Bach1 protein and hepatitis C virus gene expression in human hepatoma cells expressing hepatitis C viral proteins. *Hepatology* 51(5):1494–1504

86. Pedersen IM, Cheng G, Wieland S et al (2007) Interferon modulation of cellular microRNAs as an antiviral mechanism. *Nature* 449(7164):919–922
87. Mukherjee A, Shrivastava S, Bhanja Chowdhury J, Ray R, Ray RB (2014) Transcriptional suppression of miR-181c by hepatitis C virus enhances homeobox A1 expression. *J Virol* 88(14):7929–7940
88. Bala S, Tilahun Y, Taha O et al (2012) Increased microRNA-155 expression in the serum and peripheral monocytes in chronic HCV infection. *J Transl Med* 10:151
89. Zhang Y, Wei W, Cheng N et al (2012) Hepatitis C virus-induced up-regulation of microRNA-155 promotes hepatocarcinogenesis by activating Wnt signaling. *Hepatology* 56(5):1631–1640
90. Banaudha K, Kaliszewski M, Korolnek T et al (2011) MicroRNA silencing of tumor suppressor DLC-1 promotes efficient hepatitis C virus replication in primary human hepatocytes. *Hepatology* 53(1):53–61
91. Huang S, Xie Y, Yang P, Chen P, Zhang L (2014) HCV Core protein-induced Down-regulation of microRNA-152 promoted aberrant proliferation by regulating Wnt1 in HepG2 cells. *PLoS One* 9(1):e81730
92. Ishida H, Tatsumi T, Hosui A et al (2011) Alterations in microRNA expression profile in HCV-infected hepatoma cells: involvement of miR-491 in regulation of HCV replication via the PI3 kinase/Akt pathway. *BiochemBiophys Res Commun* 412(1):92–97
93. Nakano H, Miyazawa T, Kinoshita K, Yamada Y, Yoshida T (2010) Functional screening identifies a microRNA, miR-491 that induces apoptosis by targeting Bcl-X(L) in colorectal cancer cells. *Int J Cancer* 127(5):1072–1080
94. Hariharan M, Scaria V, Pillai B, Brahmachari SK (2005) Targets for human encoded microRNAs in HIV genes. *BiochemBiophys Res Commun* 337(4):1214–1218
95. Huang J, Wang F, Argyris E et al (2007) Cellular microRNAs contribute to HIV-1 latency in resting primary CD4+ T lymphocytes. *Nat Med* 13(10):1241–1247
96. Chiang K, Sung TL, Rice AP (2012) Regulation of cyclin T1 and HIV-1 replication by microRNAs in resting CD4+ T lymphocytes. *J Virol* 86(6):3244–3452
97. Sung T-L, Rice AP (2009) miR-198 inhibits HIV-1 gene expression and replication in monocytes and its mechanism of action appears to involve repression of Cyclin T1. *PLoSPathog* 5(1):e1000263
98. Swaminathan G, Rossi F, Sierra LJ, Gupta A, Navas-Martín S, Martín-García J (2012) A role for microRNA-155 modulation in the anti-HIV-1 effects of toll-like receptor 3 stimulation in macrophages. *PLoSPathog* 8(9):e1002937
99. Shen CJ, Jia YH, Tian RR, Ding M, Zhang C, Wang JH (2012) Translation of Pur- α is targeted by cellular miRNAs to modulate the differentiation-dependent susceptibility of monocytes to HIV-1 infection. *FASEB J* 26(11):4755–4764
100. Triboulet R, Mari B, Lin YL et al (2007) Suppression of microRNA-silencing pathway by HIV-1 during virus replication. *Science* 315(5818):1579–1582
101. Chiang K, Liu H, Rice AP (2013) miR-132 enhances HIV-1 replication. *Virology* 438(1):1–4
102. Kapoor R, Arora S, Ponia SS, Kumar B, Maddika S, Banerjee AC (2015) The miRNA miR-34a enhances HIV-1 replication by targeting PNUMS/PPP1R10, which negatively regulates HIV-1 transcriptional complex formation. *Biochem J* 470(3):293–302
103. Zhang HS, Chen XY, Wu TC, Sang WW, Ruan Z (2012) MiR-34a is involved in tat-induced HIV-1 long terminal repeat (LTR) transactivation through the SIRT1/NF κ B pathway. *FEBS Lett* 586(23):4203–4207
104. Zhang HS, Wu TC, Sang WW, Ruan Z (2012) MiR-217 is involved in tat-induced HIV-1 long terminal repeat (LTR) transactivation by down-regulation of SIRT1. *BiochimBiophys Acta* 1823(5):1017–1023
105. Chen XY, Zhang HS, Wu TC, Sang WW, Ruan Z (2013) Down-regulation of NAMPT expression by miR-182 is involved in tat-induced HIV-1 long terminal repeat (LTR) transactivation. *Int J Biochem Cell Biol* 45(2):292–298

106. Modai S, Farberov L, Herzig E, Isakov O, Hizi A, Shomron N (2019) HIV-1 infection increases microRNAs that inhibit Dicer1, HRB and HIV-EP2, thereby reducing viral replication. *PLoS One* 14(1):e0211111
107. Nathans R, Chu CY, Serquina AK, Lu CC, Cao H, Rana TM (2009) Cellular microRNA and P bodies modulate host-HIV-1 interactions. *Mol Cell* 34(6):696–709
108. Song L, Liu H, Gao S, Jiang W, Huang W (2010) Cellular microRNAs inhibit replication of the H1N1 influenza a virus in infected cells. *J Virol* 84(17):8849–8860
109. Ingle H, Kumar S, Raut AA et al (2015) The microRNA miR-485 targets host and influenza virus transcripts to regulate antiviral immunity and restrict viral replication. *Sci Signal* 8(406):ra126
110. Khongnomnan K, Makkoch J, Poomipak W, Poovorawan Y, Payungporn S (2015) Human miR-3145 inhibits influenza a viruses replication by targeting and silencing viral PB1 gene. *Exp Biol Med (Maywood)* 240(12):1630–1639
111. Wang R, Zhang YY, Lu JS et al (2017) The highly pathogenic H5N1 influenza a virus down-regulated several cellular MicroRNAs which target viral genome. *J Cell Mol Med* 21(11):3076–3086
112. Cui H, Zhang C, Zhao Z et al (2020) Identification of cellular microRNA miR-188-3p with broad-spectrum anti-influenza a virus activity. *Virology* 537(1):12
113. Ma YJ, Yang J, Fan XL et al (2012) Cellular microRNA let-7c inhibits M1 protein expression of the H1N1 influenza a virus in infected human lung epithelial cells. *J Cell Mol Med* 16(10):2539–2546
114. Chen X, Zhou L, Peng N et al (2017) MicroRNA-302a suppresses influenza a virus-stimulated interferon regulatory factor-5 expression and cytokine storm induction. *J Biol Chem* 292(52):21291–21303
115. Zhang X, Dong C, Sun X et al (2014) Induction of the cellular miR-29c by influenza virus inhibits the innate immune response through protection of A20 mRNA. *Biochem Biophys Res Commun* 450(1):755–761
116. Deng Y, Yan Y, Tan KS et al (2017) MicroRNA-146a induction during influenza H3N2 virus infection targets and regulates TRAF6 levels in human nasal epithelial cells (hNECs). *Exp Cell Res* 352(2):184–192
117. Rosenberger CM, Podyminogin RL, Diercks AH et al (2017) miR-144 attenuates the host response to influenza virus by targeting the TRAF6-IRF7 signaling axis. *PLoS Pathog* 13(4):e1006305
118. Buggele WA, Johnson KE, Horvath CM (2012) Influenza a virus infection of human respiratory cells induces primary microRNA expression. *J Biol Chem* 287(37):31027–31040
119. Hsu AC, Dua K, Starkey MR et al (2017) MicroRNA-125a and -b inhibit A20 and MAVS to promote inflammation and impair antiviral response in COPD. *JCI Insight* 2(7):e90443
120. Gui S, Chen X, Zhang M et al (2015) Mir-302c mediates influenza a virus-induced IFN β expression by targeting NF- κ B inducing kinase. *FEBS Lett* 589(24 Pt B):4112–4118
121. Zhao L, Zhu J, Zhou H et al (2015) Identification of cellular microRNA-136 as a dual regulator of RIG-I-mediated innate immunity that antagonizes H5N1 IAV replication in A549 cells. *Sci Rep* 5:14991
122. Maemura T, Fukuyama S, Kawaoka Y (2020) High levels of miR-483-3p are present in serum Exosomes upon infection of mice with highly pathogenic avian influenza virus. *Front Microbiol* 11:144
123. Wang B, Shen ZL, Gao ZD et al (2015) MiR-194, commonly repressed in colorectal cancer, suppresses tumor growth by regulating the MAP4K4/c-Jun/MDM2 signaling pathway. *Cell Cycle* 14(7):1046–1058
124. Hu Y, Jiang L, Lai W et al (2016) MicroRNA-33a disturbs influenza a virus replication by targeting ARCNI1 and inhibiting viral ribonucleoprotein activity. *J Gen Virol* 97(1):27–38
125. Dong C, Sun X, Guan Z, Zhang M, Duan M (2017) Modulation of influenza a virus replication by microRNA-9 through targeting MCP1. *J Med Virol* 89(1):41–48

126. Loveday EK, Diederich S, Pasick J, Jean F (2015) Human microRNA-24 modulates highly pathogenic avian-origin H5N1 influenza a virus infection in A549 cells by targeting secretory pathway furin. *J Gen Virol* 96(Pt 1):30–39
127. Zhang S, Li J, Li J et al (2018) Up-regulation of microRNA-203 in influenza a virus infection inhibits viral replication by targeting DR1. *Sci Rep* 8(1):6797
128. Wen W, He Z, Jing Q et al (2015) Cellular microRNA-miR-548g-3p modulates the replication of dengue virus. *J Infect* 70(6):631–640
129. Castillo JA, Castrillón JC, Dios-Toro M et al (2016) Complex interaction between dengue virus replication and expression of miRNA-133a. *BMC Infect Dis* 16:29
130. Castrillón-Betancur JC, Urcuqui-Inchima S (2007) Overexpression of miR-484 and miR-744 in Vero cells alters dengue virus replication. *Mem Inst Oswaldo Cruz* 112(4):281–291
131. Escalera-Cueto M, Medina-Martínez I, del Angel RM et al (2015) Let-7c overexpression inhibits dengue virus replication in human hepatoma Huh-7 cells. *Virus Res* 196:105–112
132. Zhu X, He Z, Hu Y et al (2014) MicroRNA-30e* suppresses dengue virus replication by promoting NF- κ B-dependent IFN production. *PLoS Negl Trop Dis* 8(8):e3088
133. Smith JL, Jeng S, McWeeney SK, Hirsch AJ (2017) A microRNA screen identifies the Wnt signaling pathway as a regulator of the interferon response during flavivirus infection. *J Virol* 91:e02388–e02316
134. Castillo JA, Castrillón JC, Dios-Toro M et al (2016) Complex interaction between dengue virus replication and expression of miRNA-133a. *BMC Infect Dis* 16:29
135. Wu N, Gao N, Fan D, Wei J, Zhang J, An J (2014) miR-223 inhibits dengue virus replication by negatively regulating the microtubule-destabilizing protein STMN1 in EAhy926 cells. *Microbes Infect* 16(11):911–922
136. Wu S, He L, Li Y et al (2013) miR-146a facilitates replication of dengue virus by dampening interferon induction by targeting TRAF6. *J Infect* 67(4):329–341
137. Kanokudom S, Vilaivan T, Wikan N, Thepparit C, Smith DR, Assavalapsakul W (2017) miR-21 promotes dengue virus serotype 2 replication in HepG2 cells. *Antivir Res* 142:169–177
138. Mitchell PS, Parkin RK, Kroh EM et al (2008) Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A* 105(30):10513–10518
139. Köberle V, Pleli T, Schmithals C et al (2013) Differential stability of cell-free circulating microRNAs: implications for their utilization as biomarkers. *PLoS One* 8(9):e75184
140. Fu X, Calin GA (2018) miR-122 and hepatocellular carcinoma: from molecular biology to therapeutics. *EBioMedicine* 37:17–18



RNA-Seq Analysis Strategies to Understand Viral Pathogenesis 11

Anvitha Nair, Arpana Vibhuti, V. Samuel Raj, and Ramendra Pati Pandey

Abstract

RNA-Seq techniques have added to associate, in treatment, growth of the knowledge distinct to the natural and cell measures needed throughout diseases. Meta-investigation of RNA-Seq tests was performed to rethink the previous proof for the higher comprehension of the pathological process. Moreover, the gene ontology and Reactome investigation performed upheld the discretionary suspicion that the antiviral pathways unremarkably moved in lightweight of intense infective agent contamination that was accountable for a debilitated growing cell. Within the meta-examination measures, it had been recognized that there are a unit 3 potential novel competition qualities inescapably enclosed throughout this antiviral response. In general, RNA-Seq may be a tremendous technique permitting depiction of cell nonuniformity among a cell community. The relationship of transcriptome identification with rapid cell aggregates might encourage clearly expressed proteins or biomarkers of interest distressed in the infectious diseases. Future specialized upgrades might conceivably beat this constraint, a minimum of halfway. Improvement of novel conventions permitting co-occurring investigation of various subatomic particles, as an example, DNA, compound, and proteins, among indistinguishable singular cell would give an extra development toward an extra careful “single-cell integrome,” which may probably be instrumental to extend our perception of uncontrollable disease.

Keywords

RNA-Seq · Pathogenesis · Transcriptome profiling · Biomarkers · Virus-host interactions

A. Nair · A. Vibhuti · V. S. Raj · R. P. Pandey (✉)
SRM University, Delhi-NCR, Rajiv Gandhi Education City, Sonapat, Haryana, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_11

11.1 Transcriptomic Analysis of Complex Viral Genomes

To start an experiment, we should design the experiments accordingly, like what methodologies and approaches should we apply before running an experiment; the primary focus should be to decide whether to work on the viral/host transcripts [1]. The second is to work for biological observations and interpretations [1]. The third is to whether document the diversity of RNAs present or to quantify the relative abundance of specific transcripts [1]. The last is to decide whether to incorporate multiple infections or to profile the early or late stages of infection based on the time points that could be optimal for the given experimental system [1]. To map transcription start site (TSS) usage or to detect alternative splicing or alternative polyadenylation, qualitative analysis can be done [1]. Based on the activity of the virus itself can make it complex to interpret the orthomyxoviruses, poxviruses, and coronaviruses, along with many herpesviruses – which either degrade or transcriptionally subdue host and viral mRNAs [1]. Lastly, it is important to take technical and analytical requirements into consideration as both are different in transcriptome studies [1]. Viruses such as herpesviruses, poxviruses, and adenoviruses, with large intense DS DNA genomes, have limitations based on their transcriptional sites, ORF, and internal splicing at 5'-3' ends [1]. Due to such complexity, the methods used to generate and interpret RNA-Seq data were appropriate for the last transcript analysis/results as it depends on the conditions of the infections and viruses [1]. The depth of sequencing may get influenced by getting a robust signal for viral RNAs [1]. In the optimization process of short-read RNA sequencing for viral transcriptomes, most of the documentation is on the host responses to different stimuli from the environment or the comparison of different cell types and tissues, with careful deliberation over the experimental design [1]. With numerous variations of the general approaches which were planned to answer specific queries like the transcript initiation plats; the placement of modified bases, etc. [1]. Standard RNA-Seq in terms of sample preparation and data analysis is quite a simple method where, depending on the experiment, isolation of complete RNA is done and later used to assemble sequenced libraries; here high-abundance rRNAs are eliminated, and the remaining is reserved in the library [1]. It is completely on the study of interest over the choice of whether non-poly(A) hosts and/or viral RNAs are to be chosen for the experiment to be performed of cDNAs, followed by simultaneous single run sequencing processes on Illumina NextGen, HiSeq, or NovaSeq platforms [1]. Later the consequent sequences are then processed to generate expression counts [1]. The genome annotations specify the limits of individually transcribed mRNAs and known splicing patterns which were generated by these expression counts, very critical to ensure the reads are conscripted to a given transcription unit, as it gets complicated as the transcription units overlap [1]. There are available annotations of genomes ranging from humans to the major model organisms, but for the viruses, it is appallingly simplified. Consequently, the alternative transcript structures easily go undetected, and in turn confounding expression level can get obscured due to the presence of overlapping transcription units leading to misinterpretations [1]. With the help of lower temperature, fragmentation time, and increased number of cycles in

the sequencing reaction, a preferred solution is acquired by altering the RNA fragmentation step (i.e., by increasing the sequence read length) [1]. A targeted enriched viral RNA offers an indispensable tool which is being incorporated into standard RNA-Seq production, where viral transcripts are scarcer than those of the host infections involved [1]. One of the major setbacks is the extensive amplification by PCR generating adequate material for sequencing, which is crucial to deduplicate accurately the mapped sequence (alignment vs reference genome/transcriptome) [1]. With the improvement in short-read RNA-Seq with selected enhancement strategies, one can conduce major unearthing in the field [1]. We already know that infections may occur through genetically identical microorganisms, with the different types of host cell responses giving rise to different after effects [1]. The attributed changes in host expression can be sensibly interpreted according to elementary concept: cellular population being homogeneous, dominated by specific cell type in the “bulk” RNA sequencing [1]. The concurrent cross-examinations of both host and microbial transcriptomes albeit tempered by some critical limitations within a single cell are becoming a reality at present, but the proposition requires a classification of the cells [1].

The incessant utilization of this suggestion encapsulates the utilization of UMI arrangements dismissing the copied succession reads from the last investigation during PCR at that point; tests pooled and eventually sequencing libraries are amassed; from the paired-end sequencing, produce one grouping read with a standardized tag, and the difference comprises a limited capacity to focus succession map at 3' closures of the mRNA [1]. The ensuing examinations focus on the ID and delineation of the host cell types by showing at least one marker; further the profiling of differential articulation is refined by spiraling rundown of devices, for example, Seurat, Monocle, MAST, etc. [1]. On account of incomplete viral genome annotations within the library, the availability of polycistronic gene arrays and the availability of non-poly(A) RNAs are leading to justifiable sequence reads being inaptly terminated or mis-apportioned, whenever 3' sequencing is conducted/performed [1]. To handle this issue, planning and resulting examinations of succession reads are finished by amalgamating host and viral reference genomes securing viral reads are all around saved and are allotted to address cell; hence, it is finished by inciting substitute viral genome comments where polycistronic quality clusters are broken down as record units [1].

Research on viral genomes such as herpesvirus and other low-abundance viral infections is a major challenge as the level of detection of viral mRNA abundance is lower in single cells [1]. We are all well aware of the advancements of the field progressing ahead in time, but the point to ponder is that experimental design and interpretation must be accurate and also the apprehensive knowledge of the prefabricated bioinformatics solutions is seldom suited for host-virus interaction analysis [1]. In the current situation, the repeat of a since quite a while ago read RNA-Seq encourages the sequencing of poly(A) mRNAs from the 3'-5' end, creating pertinently lacking reads in examination with other sequencing strategies, and has been utilized to classify record variations (through elective grafting and distinguishing novel records or record isoforms) [1]. At the point when these since

quite a while ago read RNA-Seq is joined with short-read RNA-Seq and variation draws near (CAGE-Seq), it encourages fine enumerating of viral transcriptomes at high goal [1]. To precisely plan the succession reads requires a blunder rectification that could plan the extraordinary 5' and 3' finishes of records, just to recognize locales of grafting [1]. While adjusting reads to a reference genome is moderately straightforward, these lines are regularly needed to distinguish record start and RNA cleavage destinations [1]. Information visual investigation is significant for the recognizable proof of novel qualities or joint variations [1]. Even after the upgrades in the movement of the field, it should be retrospect that, because of similarly modest number of reads produced during each run, at that point the since quite a while ago read sequencing will be impracticable if viral RNA yields are low; notwithstanding the info material would restrict the contamination models that are suitable with the predominant since a long time ago read RNA-Seq approaches [1]. Stages, for example, GridION and PromethION proffer generous number of arrangement reads per run, in spite of the fact that at the hour of drafting, nor is viable with direct RNA sequencing [1].

Bioinformaticians are these days working together with research labs (straightforwardly/in a roundabout way), basically helping in turning gathered crude information (groupings) into controlled qualities with factually huge worth, additionally called p-esteem, after conglomeratic testing [1]. Being an essential individual from the field, they distract themselves in cautious arranging of the exploratory resulting examination of the viral genome and validate no disadvantages when discussing "one (doesn't) fits all" approaches [1]. The arranging records dependent on the life form of interest they are working whereupon incorporates issues, for example, reproducibility/replication, bunch impacts, quality explanations quality and accessibility of the creature [1]. To set aside both time and cash, it is in every case better to advance the diagnostic informational collections to investigate the accepted basic issues, so there won't be any requirement for exploratory upgrade and resequencing [1]. Bioinformaticians should not incline toward benchmark RNA-Seq analysis pathways whenever handling viruses and should be apprised of the complete facet of the virus under study including biological characteristics and genome structure [1].

For example, the polycistronic clusters present restrict the scRNA-Seq quality articulation investigations in the herpesviruses in light of the fact that the total records accomplished all through the polycistronic unit allot at 3' end, which fundamentally influences the arrangement of viral reads to a transcriptome due to disposing of reads which maps indistinguishably against the 3' closures of various records, so it is compulsory to establish polycistronic qualities that allocate at 3' ends as sole record unit, which plunges the yield of organically relevant data in any case; significant natural information would wind up discarded [1]. The principle explanation for that the succession gets disposed of is that the reads got during scRNA-Seq are commonly restricted to the 3' UTR and mistakenly dispensed to a perceived quality; other planned confounder can be that the viral genomes that contains copied locales which thusly brings about either short or long grouping reads are consequently disposed of or their attitude isn't allotted precisely causing an

impact in the subsequent TPM checks; one more basic observation is that viral genomes displays in wide scope of sizes, they express the characteristics more modest than the genomes of their hosts empowering the causation of direct or roundabout genome-wide inclusion plots which gives a basic anxious outline of record designs over the viral genome [1]. Utilizing graphical yields, for example, the representation devices or through R bundles, the information can be quickly surveyed and examined against current quality comments, which could help in the disclosure of the zones inside the genome that were dark earlier and last was interpreted and maintained of elective record conformations [1].

11.1.1 RNA-Seq Evaluation

The Zika virus (ZIKV) is an arbovirus which is responsible for a range of congenital malformations including congenital Zika syndrome (CZS) that occurs during embryonic development [2]. RNA-Seq methodologies have contributed to an increase of the comprehension corresponding with biological and cellular processes involved in the course of infections [2]. Meta-analysis of RNA-Seq assays was performed to reexamine the past evidence for the better understanding of the pathogenesis of CZS [2].

In the meta-examination tests, it was distinguished that there were three potential novel competitor qualities definitely included during the antiviral reaction: APOL6, XAF1, and TNFRSF1 [2]. Eventually it got validated that IFN fading impacts the cell reaction against ZIKV infection and under basic imperative a priceless sponsor is gained by apoptotic pathways that may instigate the CZS aggregate [2]. Ordinarily revealed neurological element of CZS is microcephaly, albeit other neurological variations from the norm incorporate of brain stem brokenness, nonappearance of gulping reflex, and poly-malformation disorder; broad attributes incorporate repetitive scalp skin, anasarca, low birth weight, polyhydramnios, arthrogryposis, and ophthalmological deformities, for example, intraocular calcifications, waterfall, topsy-turvy eye sizes, macular decay, optic nerve hypoplasia, iris coloboma, and focal point subluxation [2]. As we realize that the human CNS development starts during the principal trimester of embryogenesis, here the hNPCs achieve the entirety of the glial and neuronal cell types during this period of time; the beginning of pathogenic cycles cause neuroinflammation and the discharges of the immunoregulatory particle bringing about the setting off of apoptosis, arousing a mutilation of hNPCs expansion, development, and separation; thusly imperfect mental health movement occurs [2]. There are contemplates that represents ZIKV taints hNPCs the fetal cerebrum, empowering irritation just as tissue damage [2]. Indeed, even with ongoing progressions in the portrayal of the repercussions of ZIKV disease on undeveloped CNS framework, anyway is as yet basic to distinguish which pathways are unavoidably engaged with the pathogenic cycle as this information error is unmistakably prohibitive for development of restorative methodologies which could thus fight off the extreme clinical after effect of the infection [2]. Transcriptional profiling gives an opening of correspondence between the cell capacity and

biotransformation, likewise with characterizing plausible deductions of hereditary varieties and encompassing conditions in various tissues and organisms [2].

RNA-Seq is broadly utilized over many years and has gotten the fundamental option for such investigations [2]. As per test run, they found that out of 30 examinations simply 3 were applicable to the fixed models [2]. The main investigation (Zhang et al. [SRADB id: SRP073493, GSE id: GSE8043]), it contained the chose research between the two genealogies ZIKV and dengue infection (DENV), and the hNPCs were familiar to it then among the DEG's the degrees of protein communication, transcriptional changes and quality capacity were looked at which brought about featuring 1345 DEGs among ZIKVM and "its counterfeit gathering" and 601 DEGs for the ZIKVC and the "false gathering comparison" [2]. The subsequent examination (McGrath et al. [SRADB id: SRP096367, GSE id: GSE93385]) involved the examination of the hNPC tests taken from three perished kids, from which they had the option to recognize eight up-controlled and four down-managed qualities from the common examples, when thought about by entire transcriptome profiles between the contaminated and non-tainted cells from each individual [2]. The third examination (Caires-Júnior et al. [34] [SRADB id: SRP114529, GSE id: GSE102128]) outlined the execution of RNA-Seq probe cells secured from the three sets of discrepant aggregates of CZS dizygotes; they spotted 64 DEGs (DDIT4L quality unequivocally assumes essential part in mTOR flagging pathway) [2].

For the over three investigations expressed have its own methodology (RNA-Seq information taking care of and factual strategies) subsequently they "reprocessed" their readings by directing similar convention to all the gathered examples and played out a meta-examination for it [2]. Among every one of the 17 examples, the formalization of articulation information came about around 29,318 accents present in it and distinguished 13 up-managed qualities in infection-influenced cells (rank item technique with % of bogus forecast [$p < 0.05$ and $\log_2(\text{fold change}) > 1$]), and among these outcomes, the quality metaphysics investigation perceived 847 expanded terms dependent on their foundational microorganism nature [2]. From those enlarged terms, they selected 20 terms which were conceivably connected to ZIKV contamination and reaction [2]. From the recognized 13 DEGs which came to the [$\log_2(\text{fold change})$], beginning the Reactome investigation returned 12 factually critical pathways (identified with interferon, interleukin-10, chemokines, and other receptor flagging pathways and reactions) in which the quality terms (from the quality cosmology examination) are included [2].

From all the examinations front expressed, the fundamental objective was to distinguish the potential atom that can be utilized as an antiviral palisade and furthermore to discover which particles are initiated during the antiviral reaction [2]. 2'-5'-Oligoadenylate synthetase (OAS1 and OAS3) and 2'-5'-oligoadenylate synthetase like (OASL) are connected with antiviral reactions and were discovered to be up-managing in their ZIKV-tainted supplements; these are interferon-inducible qualities which tie with dsRNA and ssRNA viral genomes and thus trigger RNase L debasement instrument of viral and cell RNA, holding onto viral production [2]. Studies implied that while ZIKV ssRNA genome is permitting the RNase L

movement, in any case, the elective situation is that the infection would effectively disregard the enzymatic action with its remarkable replication capacities in the endoplasmic reticulum [2]. The most confusing validation is portrayed by “pregulation” of interferon-inducible qualities and apolipoprotein (APOL6) qualities alongside DDX58 (RIG1) protein, which can tie dsRNA and ssRNA achieved from different flaviviruses (DENV) and elicit insusceptible reactions normally through IFN-1 creation [2].

APOL6 uncovered that besides advancing antiviral reactions against few infections, it likewise has supportive apoptotic properties [2]. There are a few vulnerabilities in the advance of expressed exploration dependent on the statement of the apoptotic proteins and furthermore in the effort of favorable apoptotic properties (for XAF1) in a critical function in the guideline of supportive provocative and resistant reactions (for TNFRSF1); however other probe shows its inclusion in the invulnerable guard component in gut mucosa; these reactions were accounted for as being associated with the ZIKV-tainted hNPCs with articulations of chemokines (CCL5, CXCL10, and CXCL11 communicated in the ZIKV-contaminated skin cells) [2].

Generally in setting with the outcomes dependent on neurogenesis, the affirmed of that the presence of ZIKV in hNPCs were detected by IFIH1 working about as a controlling viral factor in the actuation of type I interferon apoptosis and incendiary pathways with the assistance of specific cytokines and chemokines which may assume a vital part all through neurogenesis with the guideline of formative quality articulation and these indications of CZS might be because of the movement of an impending provocative response [2].

Conclusively, the profound sequencing of the complete human genome under the HGP in 2001 was the defining moment inside the “omics time.” Indeed, even late innovative advances furthermore with measurable methodology and applied science altered the affiliations with science [3, 4]. Consequently, load of actions was committed in growing new speculative chemistry, chemurgical and gadgets/machines with higher results, as Next Generation Sequencing (NGS) and Mass range investigation (MS), and propose into essential and clinical examination [5, 6]. The blend of different omics is contemporarily known because of higher general comprehension of natural cycles and sicknesses [7]. So far, a progression of novel strategies is on the lookout for test at populace level goal, with a few organic recommendations in the tempting infections [8, 9].

In record of cell non-consistency, the inside part of virology is unmistakably a worry since it is presumably going to affect infection replication cycle and subsequently setting off contamination. We realize that the ascendancy of irresistible individual replication is a lot dependent on its host, as these infections are customized in such design that would totally take resource of the cell apparatus to their own mileage.

Apparently, achieving all-inclusive contamination in an extremely given cell populace is shockingly uncommon. Non-fundamentally unrelated theories would vindicate this inconsistent status toward contamination: (1) irresistible specialist non-consistency, blend of equipped, changed, and blemished irresistible operator

particles displaying varieties in disease capacities; (2) cell non-consistency, mix of cells with varieties in digestion, creation, enactment standing, or cell cycle, following explicit cell environment molding infection movement accomplishment all through the cell. During this viewpoint, single-cell investigations typify novel open doors in distinguishing explicit cell and atomic alternatives lifting up or in differentiation prohibiting infection replication, while adding to the grip of infection have connections, coupled with furnishing new focuses to repress irresistible operator replication [10].

Standard conventions for library forecasting are upgraded to utilize exclusively some nanogram (10–100 ng) of the hereditary material. One single cell contains on normal exclusively 10 pg of absolute polymer. Likewise, work processes for polymer extraction and library arrangement were modified and upgraded to figure with single-cell material. Secluded single cells should be lysed to get to the RNA. This progression might be performed through ardent programmed gadgets or physically; or all things considered, manual cell lysis and complete polymer filtration might be performed correspondingly misusing sap-based sections (e.g., single-cell RNA purification kit (NorGen Biotek Corp), PicoPure1RNA isolation kit (Thermo Scientific)) or attractive dab partition (e.g., 5-min single-cell polymer extraction kit (Biofactories)). Following this, the polymer is advanced for mRNA, either by poly (A) decision or at times by cell organ consumption [11, 12].

The enhanced RNA portion later is expectedly opposite interpreted with altered oligo-dT groundwork through conflicting conventions, either by poly(A) following or model change. During the RT step, a few conventions allow labeling of single atoms with contrastive subatomic identifiers (UMI); irregular hexanucleotides might be acclimated and measure unequivocally the amount of starting mRNA particles that might be available in one single cell [13, 14].

Succeeding converse record, cDNA gets enhanced by in vitro record or through PCR. The enhanced cDNA library is then utilized for record mixture and high-output sequencing for additional information on RNA-Seq methodology need to look for rules from single-cell information investigation [15, 16].

As each single cell is selective, it's unsuitable to execute investigative deception and evaluate disorder. It's along these lines fundamental to claim some qc to ensure information steadfastness. This could be obtained by adding false mRNAs of known grouping and amount, similar to external polymer controls partner (ERCC) polymer spike-ins, to every cell lysate. The amount of reads rediscovered from the spike-ins can offer information concerning between test-specialized inconstancy [17].

It has been as of late demonstrated that sequencing bumbles in UMI succession are conventional and will subsequently bias record evaluation whenever utilized. To devalue the mistake pace of UMI all through sequencing. Lau et al. built up a substitute methodology in developing a high error-safe dramatically extended scanner tags (EXBs) with giga-scale variety and familiarize them into ds-cDNA by methods for a transposase [18].

References

1. Depledge DP, Mohr I, Wilson AC (2018 Dec 10) Going the distance: optimizing RNA-Seq strategies for Transcriptomic analysis of complex viral genomes. *J Virol* 93(1):e01342–e01318
2. Gratton R, Tricarico PM, Agrelli A, Colaço da Silva HV, Coêlho Bernardo L, Crovella S, Campos Coelho AV, Rodrigues de Moura R, Cavalcanti Brandão LA (2020 Feb 17) In vitro zika virus infection of human neural progenitor cells: meta-analysis of RNA-seq assays. *Microorganisms* 8(2):270
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al (2001) The sequence of the human genome. *Science (New York, NY)* 291:1304–1351
4. Manzoni C, Kia DA, Vandrovцова J, Hardy J, Wood NW, Lewis PA, Ferrari R (2016) Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief Bioinform*:1–17. <https://doi.org/10.1093/bib/bbw114>
5. Haque A, Engel J, Teichmann SA, Lonnberg T (2017) A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 9:75
6. Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W, Shi L (2011) Nextgeneration sequencing and its applications in molecular diagnostics. *Exp Rev Mol Diagn* 11:333–343
7. Ohashi H, Hasegawa M, Wakimoto K, Miyamoto-Sato E (2015) Nextgeneration technologies for multiomics approaches including interactome sequencing. *Biomed Res Int* 2015:104209
8. Tanca A, Addis MF, Uzzau S (2013) High throughput genomic and proteomic technologies in the fight against infectious diseases. *J Infect Dev Ctries* 7:182–190
9. Jean Beltran PM, Federspiel JD, Sheng X, Cristea IM (2017) Proteomics and integrative omic approaches for understanding host– pathogen interactions and infectious diseases. *Mol Syst Biol* 13:922
10. Rato S, Golumbeanu M, Telenti A, Ciuffi A (2017) Exploring viral infection using single-cell sequencing. *Virus Res* 239:5568
11. Fan X, Zhang X, Wu X, Guo H, Hu Y, Tang F, Huang Y (2015) Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol* 16:148
12. Sheng K, Cao W, Niu Y, Deng Q, Zong C (2017) Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat Methods* 14:267–270
13. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lonnerberg P, Linnarsson S (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 11:163166
14. Smith T, Heger A, Sudbery I (2017) UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res* 27:491–499
15. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA (2015) The technology and biology of single-cell RNA sequencing. *Mol Cell* 58:610–620. Nice and comprehensive review about the whole scRNA-Seq pipeline.
16. Picelli S (2017) Single-cell RNA-sequencing: the future of genome biology is now. *RNA Biol* 14:637–650. Description of the diverse methods used to generate the scRNA-Seq libraries.
17. Stoeger T, Battich N, Pelkmans L (2016) Passive noise filtering by cellular compartmentalization. *Cell* 164:1151–1161
18. Lau BT, Ji HP (2017) Single molecule counting and assessment of random molecular tagging errors with transposable gigascale error-correcting barcodes. *BMC Genomics* 18:745



Various Transcriptomic Approaches and Their Applications to Study Small Noncoding RNAs in Dengue and Other Viruses

12

Deeksha Madhry, Kush Kumar Pandey, Shivani Malvankar, Shubham Kumar, Anjali Singh, Ravi Kumar S. Yelegara, Rupesh K. Srivastava, and Bhupendra Verma

Abstract

Advanced deep sequencing technologies have revolutionized our understanding toward noncoding RNAs (ncRNAs) and have uncovered their various regulatory roles which are performed by the fine-tuning of host gene expression at either epigenetic or transcriptional and posttranscriptional level. Rapid development in various deep sequencing technologies and bioinformatics platforms has targeted ncRNAs for therapeutic purposes. Here we are summarizing various transcriptomic techniques, bioinformatics tools, and databases and their application to understand the modulation of various regulatory ncRNAs in context of dengue and other viral pathophysiology.

Keywords

Noncoding RNAs · Dengue virus · EST · Next generation sequencing · Illumina sequencing · Ion Torrent sequencing · Pyrosequencing · Oxford Nanopore · PacBio sequencing

Abbreviations

CAGE Cap analysis of gene expression
EST Expressed sequence tag
lncRNA Long noncoding RNA

D. Madhry · K. K. Pandey · S. Malvankar · S. Kumar · A. Singh · R. K. Srivastava · B. Verma (✉)
Department of Biotechnology, All India Institute of Medical Sciences, New Delhi, India
e-mail: bverma@aiims.edu

R. K. S. Yelegara
Department of Biotechnology, M. S. Ramaiah Institute of Technology, Bengaluru, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_12

195

MPSS	Massively parallel signature sequencing
SAGE	Serial analysis of gene expression
seq	Sequencing
sncRNA	Small noncoding RNA

12.1 Introduction

Viral infections often lead to the modulation of various regulatory RNAs like miRNA, siRNA, tRNA-derived RNA fragments, Piwi RNA, etc. that are called noncoding RNA. These ncRNAs do not code for any protein. Modulation of small ncRNAs (sncRNAs) during viral infection is primarily associated with the regulation of endogenous host gene expression, and expression of some endogenous virus generated noncoding RNA in host cells which may potentially contribute to disease progression. vsRNA and sfRNA are virus-derived small noncoding RNAs which are found to be associated with the evasion of host antiviral response. Table 12.1 depicts various host- and virus-originated noncoding RNAs and their roles during viral infections.

Delayed diagnosis of infectious disease often aggravates the disease condition. Differential expression of sncRNA can prove to be dispensable biomarkers to identify the disease severity during various stages of infection. RNA-based therapeutics represents next-generation sequencing approaches for prospective antiviral and other disease-related therapeutic development. Many RNA therapeutics have already been approved by the FDA such as miravirsin to treat HCV infection and patisiran, givosiran, and MRX34 to treat various tumors. Thus unraveling the transcriptional landscape of various infectious disease is important to build an articulated nexus of regulatory roles of ncRNA during virus-host interaction. Recent advancements in the field of deep sequencing and high-throughput screening have enabled to establish a deeper understanding about RNAome and thus explore the capacities of these sncRNAs. Here we are summarizing various transcriptomic techniques used to study dengue virus and other viruses to study small noncoding RNAs.

12.2 Tag-Based Approaches

The tag-based approaches have been used for the direct determination of the cDNA sequences. These include expressed sequence tag (EST), serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE), and massively parallel signature sequencing (MPSS). The expressed sequence tag has proved to be a rapid and efficient means of characterizing the huge sets of gene sequences consisting of 300–1000 bp of DNA, and it is often accumulated in a database as a “single-pass read” that is sufficient in establishing the identity of particular expressed gene. Serial analysis of gene expression (SAGE) is another method used to obtain the qualitative and quantitative level of gene expression profiles in different conditions. The SAGE methodology is considered as an “open architecture” technique, and unlike another

Table 12.1 Types of noncoding RNAs with functions and examples in virus infections

Noncoding RNA types	Long name	Functions	Examples	References
Small ncRNAs	miRNA (19–24 bp)	mRNA degradation, transcriptional and translational repression	Proviral miR-146a DENV, herpesvirus infection Antiviral miR-133a in DENV infection	[1, 2]
	piRNA (24–31 bp)	Transcriptional and transposon silencing, transgenerational inheritance, germline cell viability	Regulation of viral replication in human papillomavirus and rift valley fever virus	[3, 4]
	tRFs	mRNA stability, inhibition of translation initiation or elongation, signaling molecule, regulation of apoptosis, ribosome biogenesis, epigenetic factor	Proviral tRFs in respiratory syncytial virus infection Antiviral tRFs in HIV infection	[5, 6]
Mid-sized ncRNA	snoRNAs	Posttranscriptional modification, maturation of rRNA, small nuclear RNA, and other cellular RNA	SNORD3 promotes pre-rRNA formation SNORD115 plays a role in alternative splicing SNORD28 precursor for miRNA	[7, 8]
	PASRs	Epigenetic modification that affects gene transcription	Regulate gene expression by binding with AGO complex	[9]
	TSSa-RNAs	Transcriptional start site-associated RNA	Associated with RNF12 and CCDS2 genes	[10]
	PROMPTs	Promoter upstream transcripts (<200 bp)	HIF2PUT, a PROMPT for HIF2 α , acts as inhibitor of osteosarcoma stem cells	[11, 12]

(continued)

Table 12.1 (continued)

Noncoding RNA types	Long name	Functions	Examples	References
Long-sized ncRNA	lncRNA Long noncoding RNA (>200 bp)	Epigenetic modification, transcriptional and posttranscriptional regulation	lncATV required for Zika virus replication	[13]
	T-UCRs Transcribed-ultraconserved region (>200 bp)	Regulation of gene expression by directly interacting with mRNA or by interacting with miRNA	uc283 expressed in cancer cells uc38 a tumor suppressor in breast cancer	[13, 14]
Others	eRNA Enhancer RNA (50–2000 bp)	Activation of transcription	IL1 β -eRNA required for expression of IL1 β which has a role in inflammatory responses	[15]
	circRNA Circular RNA (>200 bp)	Act as miRNA sponge, cell proliferation, regulation of transcriptional rate; regulate function of RNA binding proteins	CIRS-7 inhibits miRNA-7 which improves insulin secretion in diabetes circRNA generated during Kaposi's sarcoma herpesvirus shows antiviral response	[16, 17]
Virus-derived ncRNA	PARs Promoter-associated RNAs (16–200 bp)	Transcriptional regulation	PARs regulate gene expression in malignant melanoma	[17]
	sfRNA Sub-genomic flavivirus RNA (300–500 bp)	Viral transmission and pathogenesis	sfRNA role in transmission of Zika virus from <i>Aedes</i> mosquito West Nile virus sfRNA evades host interferon antiviral response	[18, 19]
	vsRNA Viral small RNA (23 bp)	Regulation of viral replication	HHS-6 snRNA-U14 is abundantly expressed in herpes infection required for activation of viral replication	[20]

array method, it does not need prior information about the genes which needs to be analyzed, and it also reflects the absolute mRNA levels. Further, several SAGE-like methods have also been developed. They can be employed for the genome-wide analysis of DNA copy-number changes, transcription factor targets, and analyzing epigenetic signatures such as methylation patterns and chromatin structure. CAGE is another high-throughput sequencing-based technique that enables us to quantify and identify the expression of 5' capped RNAs [21]. Likewise, MPSS is an open-ended platform for quantifying in-depth gene expression based on individual mRNA levels in the cell. In MPSS, the identification and characterization of the gene prior to conducting the experiment are not required. It has the routine sensitivity at the level of mRNA molecules, and the datasets are in the digital format that facilitates the management and analysis of data.

12.3 Chip-Based Approaches

Microarray plays a significant role in transcriptome profiling studies. Microarray is used for DNA mapping, sequencing, and transcript-level analysis [22]. It's a crucial genomic technology and often used synonymously with DNA microarray and high-throughput gene expression analysis. Gene expression microarray is a nucleic-acid-hybridization-based technique, and a complex mechanism is required to understand the global and parallel analysis of different cellular processes. It's been more than a decade that microarray has been in use as a gold standard for transcriptome studies in a wide range of settings [23]. Recently, despite the emergence of the next-generation sequencing [24], DNA microarrays still appear an extremely compelling approach due to its quick, precise, and inexpensive detection of the pathogens compared to culture or immunoassay techniques [25, 26]. It is a technique of choice for the detection of altered gene expression upon virus infection and is often used in clinical diagnostic to detect for the presence of existing viruses or new viruses [27, 28].

Tiling arrays are subtypes of microarray which are used to investigate thousands of gene expression in a coordinated fashion, transcriptome mapping, and to identify the transcription factor binding sites [29, 30]. It is a useful tool for the investigation of whole genome or chromosome expression as well as to uncover the various novel RNA expression patterns [31, 32]. Like traditional oligonucleotide microarrays, the probes are designed to the matched parts of the genomic region of interest. The probes get hybridized to the labeled DNA or RNA target molecules which are fixed onto a chip. The experimental technique to identify the site for protein DNA interaction involves the hybridization of immunoprecipitated DNA on a tiling microarray (ChIP-chip experiments) [33]. The genome-tiling microarrays have facilitated the analysis of global expression patterns with or without completely annotated genome in organisms such as prokaryotes, mouse, human, and yeast [34, 35]. In conventional gene-probing microarrays, probes targeting a particular gene give independent measures of the same RNA expression, while when tiling is the strategy applied to the entire genome, the analysis becomes restricted to the annotated genes, and it becomes difficult for the unannotated genes to be analyzed.

Therefore, major challenges for tiling array studies are (1) determining transcriptional start and stop sites and (2) predicting whether transcripts are either from long continuous stretches such as genes or from short noncontinuous strands of RNA, i.e., typically ncRNAs [36, 37].

Although microarray-based transcriptional profiling was very much prevalent during the last decade and have made numerous remarkable contributions, with the emergence of new sequencing technologies, its use has become limited. Although there is an advanced microarray application in regard to RNA profiling, however, it has certain limitations in terms of reproducibility, sensitivity, and specificity [38].

12.4 RNA Sequencing Technologies (Platforms)

12.4.1 RNA Sequencing

RNA seq is a recently established high-throughput deep sequencing technique which has been used extensively for mapping, quantifying, and identification of novel genes. Small noncoding RNAs has been widely studied these days since these have been found to play a critical role in regulation of gene expression. After microarray, RNA sequencing is slowly procuring its space in the profiling studies and now has become the gold standard for a novel small RNA discovery as well as for small RNA profiling with high throughput, high sensitivity, and reproducibility. Deep RNA sequencing does not need any prerequisite information about the sequence which is a clear advantage over existing approaches like microarrays and qPCR.

12.4.1.1 Workflow

RNA seq workflow typically comprises of small RNA enrichment, RNA assessment and quantitation, small RNA library construction, deep sequencing, and bioinformatics analysis. Figure 12.1 is showing a typical RNA seq workflow.

12.4.1.2 Small RNA Enrichment Methods

Small noncoding RNAs play a significant role in gene expression. In order to analyze them, additional purification steps are required as they are present in very low concentration and thus enrichment methods are employed for their isolation and purification. Two types of enrichment methods can be used, i.e., manual methods and kit-based approaches.

Manual Methods for Isolation of sncRNAs

miRICH Method

miRICH method is variant of TRIzol method used for total RNA isolation. In this method, total RNA is isolated by using TRIzol reagent, and in order to concentrate small RNAs, washing using ethanol is skipped, after the precipitation step and

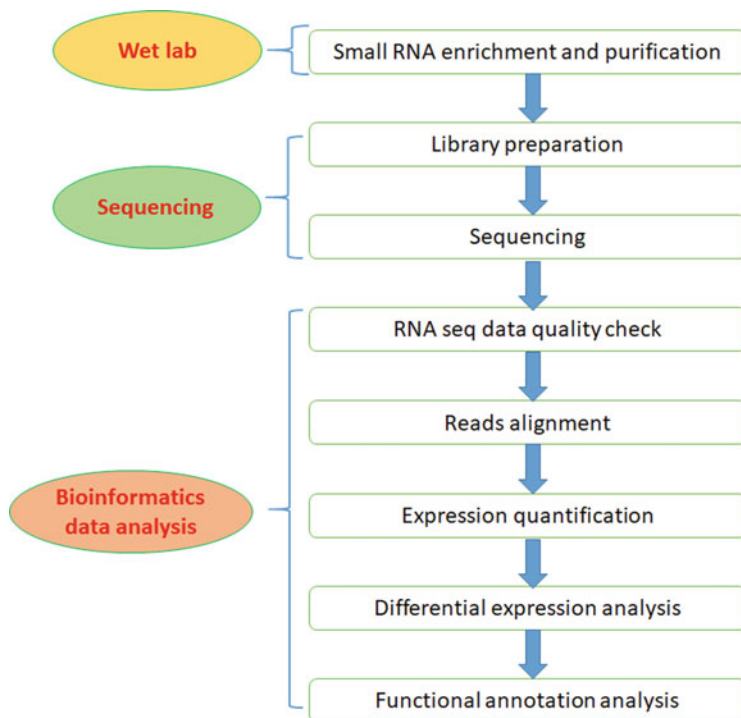


Fig. 12.1 Workflow of a typical RNA sequencing experiment and analysis

the pellet is overdried. Overdrying the pellet causes aggregation of larger RNA molecules due to high salt concentration; hence larger molecules are barely soluble, while small RNAs can be easily eluted [39].

Fractionation Approach

Fractionation method makes use of flashPAGE fractionator that works on the principle of gel electrophoresis. sncRNAs are isolated based on their size on a miniature PAGE. A precasted PAGE is inserted into the buffer chamber of the fractionator after addition of lower running buffer; this is followed by addition of upper running buffer. The small RNA fraction will be present in lower running buffer after electrophoresis which further enriched by using flashPAGE cleanup kit [40].

Multidimensional HPLC for Purification of Small Noncoding RNA

mRNA isolation can be eased by exploiting poly(A) tail, while isolation of sncRNA is difficult because of limited knowledge regarding biochemistry, posttranscriptional processing, ribonucleotide modification, and biological function, hence limiting the ability to obtain RNA in pure form. HPLC can be used to purify all classes of ncRNA

from single samples. To achieve complete separation of two ranges, size exclusion chromatography with ion-pair reverse phase chromatography can be used [40].

p19 Protein-Mediated Isolation of siRNA and Detection of miRNA

Another approach for enrichment of sncRNA is to make use of the property of p19 protein to strongly bind to siRNA derived from plant virus, Carnation Italian ringspot virus (CIRV). The p19 fusion protein designed can be used for isolation of siRNA and detection of miRNA with the help of magnetic beads [41].

Trans-Kingdom Rapid Affordable Purification of RISC (TRAP) for Isolation of all Classes of Silencing Small Noncoding RNA

Isolation of small noncoding RNAs from total RNA using size separation methods is tedious and may give poor yields. An alternative to total RNA isolation method is the AGO purification technique in which the AGO-associated sncRNA complex is purified in order to isolate small RNA by anion exchange chromatography. The issue of contamination by other RNA species will be overcome by employing this technique, but it requires highly specific antibody against AGO proteins. Anion exchange chromatography makes use of Q Sepharose a positively charged anion exchange matrix to adsorb the RNPs complex, and then the complex is eluted by increasing mild salt concentration. Further analysis is based on the gel electrophoresis technique [42].

Kit-Based Methods

Silica-based columns are used for recovery of sncRNAs widely. Different commercial kits are available for isolation of small RNAs like mirVana, miRNeasy mini kit, mirPremier, MasterPure RNA purification kit, and miRCURY as described in Table 12.2. Selection of method depends upon the initial volume of sample, type of sncRNA to be analyzed, ease to use, and price per sample [42].

Table 12.2 Various kits available for small RNA enrichment

Method	Description
mirVana kit	Uses glass fiber filters for isolation of sncRNA after organic extraction method. In presences of higher concentration of ethanol, small RNAs bind to the column while larger are eluted out
miRNeasy	Is similar to mirVana kit which employs both organic extraction and solid-based extraction of small RNAs. Silica column is used and the technique can be automated using QIAcube Connect
mirPremier kit	Biological samples are lysed using lysis mixture which releases small RNAs and genomic DNA, and large RNA molecules remain insoluble and are removed along with cell debris by short centrifugation
Every RNA purification system	Isolation of small RNAs from extracellular vesicles using this kit is possible by its ability to capture total RNA including small RNAs from extracellular vesicles

12.4.1.3 RNA Quality and Quantity Assessment

For assessment of RNA quality, two methods are used, i.e., denaturing/non-denaturing agarose gel and capillary electrophoresis, which determine intact RNA by identifying 28S and 18S rRNA bands [43]. Quantity of RNA can also be determined by various methods. It includes UV absorbance at wavelength 260 nm, fluorometry which employs fluorescently labeled probe specific to sncRNAs which have to be quantified, and splinted ligation assay where a small RNA-specific bridge oligonucleotide is used to form base pairs with the 5'-end-radiolabeled oligonucleotide and sncRNA [44]. For real-time quantification of small RNA, stem loop PCR is used in order to increase the length of template, and quantification is done using TaqMan probe [45].

Recent development in microfluidic technology has led to development of different instruments which can check RNA integrity as well as quantified RNAs. For example, microfluidic instruments like Agilent 2100 Bioanalyzer, 2200 TapeStation, and Experion can check RNA integrity by calculating 28S /18S ratio, and quantification is done by fluorescent labeling.

12.4.1.4 Small RNA Library Construction

Small RNA library preparation typically starts with the ligation of DNA adapters. It includes ligation of pre-adenylated adapter oligo at the 3' end of the sncRNA using a truncated variant of T4 RNA ligase2. It is followed by ligation of adapter oligo at the 5' end using T4 RNA ligase 1. The sequence is then reverse transcribed into cDNA following ligation. Although direct RNA sequencing is possible, but many instruments are based upon DNA sequencing due to which cDNA library preparation is done and is the crucial step of sequencing. After cDNA preparation, PCR amplification is carried out. For the longer sequences, fragmentation is done using DNase I treatment or sonication before adaptor ligation. But small RNAs can be sequenced directly after adapter ligation since these already exists as a shorter sequence. The adaptor ligation results in loss of strand specificity which is restored by dUTP incorporation in the second strand of cDNA which is then further degraded by uracil-DNA glycosylase (UDG) before the amplification step. Hence, only a single strand is amplified by PCR up to 8–12 cycles. Finally, single-end sequencing (from one end) or paired-end sequencing (from both ends) can be carried out in a high-throughput manner. The read sequences obtained are then aligned with the reference genome present in the database; else de novo assembly is carried out for the reads.

12.4.1.5 Bioinformatics Analysis

Once the raw sequencing data is generated, it is then preprocessed and normalized which involve trimming of adaptors, alignment, outlier removal, RNA class filtering, and generation of unique reads. Normalization involves comparison of expression levels across libraries. There are two commonly used small RNA database: (1) miRBase comprises information of all miRNA sequences and annotations; (2) Rfam is an open-access database which includes information about tRNA, rRNA, snoRNA, etc.

12.4.1.6 Challenges

Library Preparation Challenges

RNA seq contains various manipulation stages during cDNA library preparation, which can complicate the profiling procedure of the transcripts. Sometimes, sequencing data is incomparable with the qPCR, Northern blot, and microarray results which predominantly can happen due to the bias that occurs during different library preparation steps. Many identical short reads are amplified from cDNA libraries which give a reflection of abundant RNA species and thus culminate in false results.

Small RNA ligation bias is another attribute for library preparation which is particularly dependent upon T4 RNA ligase during adaptor ligation. It has been suggested that randomization of the adapter sequences near ligation junction might result in bias reduction. Illumina sequencing identifies antisense as well as overlapping transcripts which are unwanted. To avoid this, dUTP is incorporated during the generation of second cDNA synthesis prior to adaption ligation. The strand containing dUTP then finally gets digested leaving only a single strand for amplification.

Bioinformatics challenges include storage, retrieval, and processing of large amount of data which are very much essential for error reduction during result analysis. Background noise also occurs due to incomplete processing of RNAs.

12.4.2 Second-Generation Sequencing Platforms.

12.4.2.1 Pyrosequencing

This technique is based on the principle of sequence by synthesis approach which detects released pyrophosphate (PPi) during DNA synthesis. To detect the nucleic acid synthesis, four enzymatic reactions take place in a sequential manner. The primer is first hybridized with the biotin-labeled single-stranded DNA template. The hybridized primer template is then mixed with the enzymes: DNA polymerase, ATP sulfurylase, luciferase, and apyrase, and the substrates adenosine 5'-phosphosulfate (APS) and luciferin. Each dNTP is added separately to the reaction mixture. With the incorporation of a single nucleotide, one PPi is released which is then quantified in an equimolar concentration to the incorporated nucleotide. ATP sulfurylase forms ATP from PPi in the presence of APS. ATP and luciferase catalyze the conversion of luciferin to oxyluciferin which produces visible light which is accountable for the total amount of ATP generated. The produced light is detected by a photon detection device, with a maximum wavelength of 560 nm. ATP and unincorporated dNTPs are continuously degraded by apyrase [46].

Currently, pyrosequencing has been widely used in single nucleotide polymorphism (SNP) genotyping identification of bacteria, fungal, and viral typing. Moreover, this method can be used for the determination of difficult secondary structures and for the identification of mutations. Other possible applications are DNA methylation analysis and whole genome sequencing [47].

12.4.2.2 Ion Torrent Sequencing

Ion Torrent also requires sequence amplification, but it is the first technique that uses electrochemical detection, not camera scanning and fluorescence detection. Ion Torrent is done with an Ion Personal Genome Machine (PGM). Firstly, the DNA sample is prepared from fragmented RNA and then linked together. The library is then clonally amplified by emulsion PCR onto beads known as Ion Sphere particles. The beads are placed into proton-sensing wells on a semiconductor sequencing chip so that one bead can roughly fit into hundreds of millions of wells. The chip is then submerged into a nucleotide solution and results in the release of protons and a corresponding change in pH. This change in pH is documented by the PGM to determine whether the right nucleotide was used in the process or not, let alone if a nucleotide was added at all. As sequencing occurs, each of the four bases is introduced sequentially. A clear indication of the correct nucleotide being added is the presence of voltage. There will be no voltage found if the wrong nucleotide is added, and there will be double the voltage if two nucleotides are added. The more nucleotides present, the greater the increase in voltage and pH [48].

This technique can be used for targeted DNA and RNA sequencing, exome sequencing, viral typing, bacterial typing, aneuploidy and CNV analysis, and small RNA and miRNA sequencing [49].

12.4.2.3 Illumina Sequencing

It is often known as Solexa sequencing based on the name of the founder Solexa. It is based upon sequencing by synthesis approach and reversible dye terminators that are fluorescently labeled. These terminators are incorporated to the new DNA strand and stop DNA synthesis. These are then detected by the camera. Each type of dNTP is differently fluorophore-labeled, and therefore each base addition gives off different light which is detected by laser [50].

It is the next-generation sequencing that is used for the small RNA discovery, metagenomics, methylation profiling, transcriptome analysis, genome-wide profiling, and protein-RNA/DNA interaction analysis [51].

12.4.2.4 ABI SOLiD Sequencing

This principle is based upon the principle of sequencing by ligation. PCR-amplified target sequence is anchored by agarose beads onto a glass surface. Fluorescently labeled oligonucleotides are then added for sequential ligation mediated by enzyme DNA ligase. Once labeled oligonucleotides are annealed; fluorescent gets removed from the fragment due to the formation of phosphonothioate bond between the bases. The removal of fluorophore from 5' site makes that site vacant for upcoming nucleotide ligation. This allows the formation of different fluorescent peaks corresponding to different nucleotides [52].

It is used for targeted sequencing, identification of small RNAs, epigenome analysis, and chromatin- immunoprecipitation etc. [53].

12.4.3 Third-Generation Sequencing Platforms

12.4.3.1 Oxford Nanopore Sequencing

Oxford sequencing is based upon the principle of small variations in electrical conductance that is generated when a DNA strand or a nucleotide passes through a pore of 1.5 nm size. The nanopore is immersed in a conducting fluid along with tiny wells layered with DNA polymerase which catches the hold of the exposed DNA and makes it pass through the nanopore. Each nucleotide passage through the pore creates varying degrees of obstruction that leads to the generation of varying electric current. This small variation in the electric current defines the characteristic of each nucleotide. This change of current is recorded to detect the sequence [54].

This technique confers precision and sensitivity and is cost-effective, and also, RNA can be directly sequenced with this technique. Very long 10–50-kb-sized fragments can also be sequenced. It can be used for metagenomics analysis, de novo assembly, sequencing of long transcripts, etc. [55].

12.4.3.2 PacBio Sequencing

PacBio is again based upon the principle of sequencing by synthesis approach which makes use of specialized SMRT chips for sequencing. All the four bases are fluorescently labeled with different fluorophores. Upon binding of the polymerase to the template on the SMRT chip, each incorporating base will be irradiated with different fluorescent colors. The peak and wavelength of the fluorescent emission are then recorded for the sequencing [11, 12].

Unlike first- and second-generation sequencing, this technique requires no PCR amplification and can sequence even longer reads. This technique is used for de novo sequencing of genomes and transcriptomes, detecting alternative splicing isoforms, epigenetic modifications, analysis of different mutations, etc. [56] (Table 12.3).

12.5 Data Processing and Analysis

Next-generation sequencing (NGS) technologies have been emerging with the potential to explore small RNA (sRNA) transcriptomes and their associated roles. It offers a reliable and high-throughput approach for the identification and quantification of various sRNA classes. The steps required for RNA sequence analysis of sncRNAs has been summarized in Fig 12.2. Various tools are available online for the analysis of a particular small RNA from sequencing data, shown in Table 12.4. Various tools are also available for the integrated analysis of more than type of RNA analysis which are depicted in Table 12.5.

12.6 Available Databases for Studying Small Noncoding RNAs

1. *miRBase*: It consists of the miRNA database with information about sequence, genomic location, and predicted targets of miRNA [78].

Table 12.3 Advantages and disadvantages of different transcriptomic approaches

Technique	Advantages	Disadvantage
Microarray	<ul style="list-style-type: none"> • Well-defined hybridization protocol and analysis pipeline • Standardized approaches to submit the data and relatively cost-effective • Data resolution dependent on genome size >35–40 kb for human and mouse • Works for more than 1000 genes simultaneously • Ready to use commercial array chips are available 	<ul style="list-style-type: none"> • Analysis of predefined sequences with limited scanner dynamic range • Relies on hybridization which is potentially nonspecific • High variance for low expressed gene and may not give the paralogue information and not useful in identifying splice variants • Detection of SNP mutations is limited • Requires competency for data normalization and analysis
Tiling array	<ul style="list-style-type: none"> • Can identify and quantify up to hundred-fold expression level of transcripts • Complete coverage of genome • New splice forms can be discovered • Sufficient information for quantitative determination can be obtained 	<ul style="list-style-type: none"> • Cannot be used to analyze various isoforms and allelic expression • High background noise • Requires large number of probes • Difficult in analysis • Probe characteristics are largely variable • Chances of cross- hybridization • Determining transcriptional start and stop sites
SAGE	<ul style="list-style-type: none"> • Allows identification of novel transcripts • Can directly estimate gene abundance level with absolute data • Capable to distinguish different isoforms and allelic expression • High-throughput method and does not require any special devices 	<ul style="list-style-type: none"> • SAGE library preparation is a quite complex step • Relatively, it has low throughput due to the lack of anchoring enzyme site • Poor identification of unannotated transcripts
CAGE	<ul style="list-style-type: none"> • Useful in identifying the promoter region and transcription start site 	<ul style="list-style-type: none"> • Limited to the 5' capped transcripts
EST	<ul style="list-style-type: none"> • Highly informational content • Excellent method in providing qualitative alterations in gene expression • Allows for the identification and analysis of precise gene polymorphisms and mutations 	<ul style="list-style-type: none"> • cDNA library preparation is quite complex step • Wide transcriptome profiling requires higher sequencing cost • More amount of RNA is required • Low-throughput method
RNA sequencing	<ul style="list-style-type: none"> • Discovery of novel RNA species and RNA profiling without any prior knowledge of sequence • Offers higher sensitivity and a broader range of expression • Identify RNA splicing detection 	<ul style="list-style-type: none"> • Amplification of identical short reads gives false results • Fragmentation creates bias in the outcome • Loss of strand specificity of the library
ABI SOLiD	<ul style="list-style-type: none"> • Higher accuracy of sequencing data 	<ul style="list-style-type: none"> • Low throughput • Shorter read length • Time-intensive

(continued)

Table 12.3 (continued)

Technique	Advantages	Disadvantage
Pyrosequencing	<ul style="list-style-type: none"> • Fast method with real-time read-out • Sample preparation is relatively quick than Sanger sequencing • Reagents are also lower in cost 	<ul style="list-style-type: none"> • Problem with de novo sequencing of polymorphic regions in heterozygous DNA material • Difficulty in determining the incorporated nucleotides in homopolymeric regions
Ion Torrent	<ul style="list-style-type: none"> • With quick turnover rates, limited quantitative data, and small instrument size • Allows for many more reads to be done per sequencing run • Instrument upgraded through disposable chips • Less complex machine • Clear trajectory to improve performance 	<ul style="list-style-type: none"> • Higher error rate than Illumina
Solexa (Illumina) sequencing	<ul style="list-style-type: none"> • High-throughput DNA sequencing in a limited time • Possible to carry out sequencing of 1 terabase (TB) data per day • High accuracy up to 99.9% • Better performance in the sequencing of homopolymeric regions as compared to other sequencing techniques 	<ul style="list-style-type: none"> • Substitution errors are observed more commonly due to background noise at each cycle of sequencing • Scars persist on nucleotide structure after cleavage of blocking group which eventually causes decreased efficiency of sequencing reactions
Oxford Nanopore sequencing	<ul style="list-style-type: none"> • Real-time sequencing of single molecules at low cost • High throughput • No amplification required 	<ul style="list-style-type: none"> • Have some temperature limitations • Controlling the speed of ssDNA passing through nanopore is yet to be achieved at maximum frequency • High error rates >4%
PacBio	<ul style="list-style-type: none"> • Requires no PCR amplification • Covers sequences with GC-rich and high-repeat region • Most reliable to quantify low-frequency mutation • Provides very long reads with average read length of 8–15 kb and up to 40–70 kb • Time-effective at the rate of 10 nt per second 	<ul style="list-style-type: none"> • High error rate (around 11–15%) • Relatively low throughput

2. *MirZ*: It consists of the miRNA database with information about sequence-based miRNA profiles and predicted targets [79].
3. *IsomiR database*: It mainly has information about miRNAs and isomiRs with respect to the reads; isomiRs assigned to miRNAs belong to human 293 T cells, with miRNA annotation from miRBase [80].

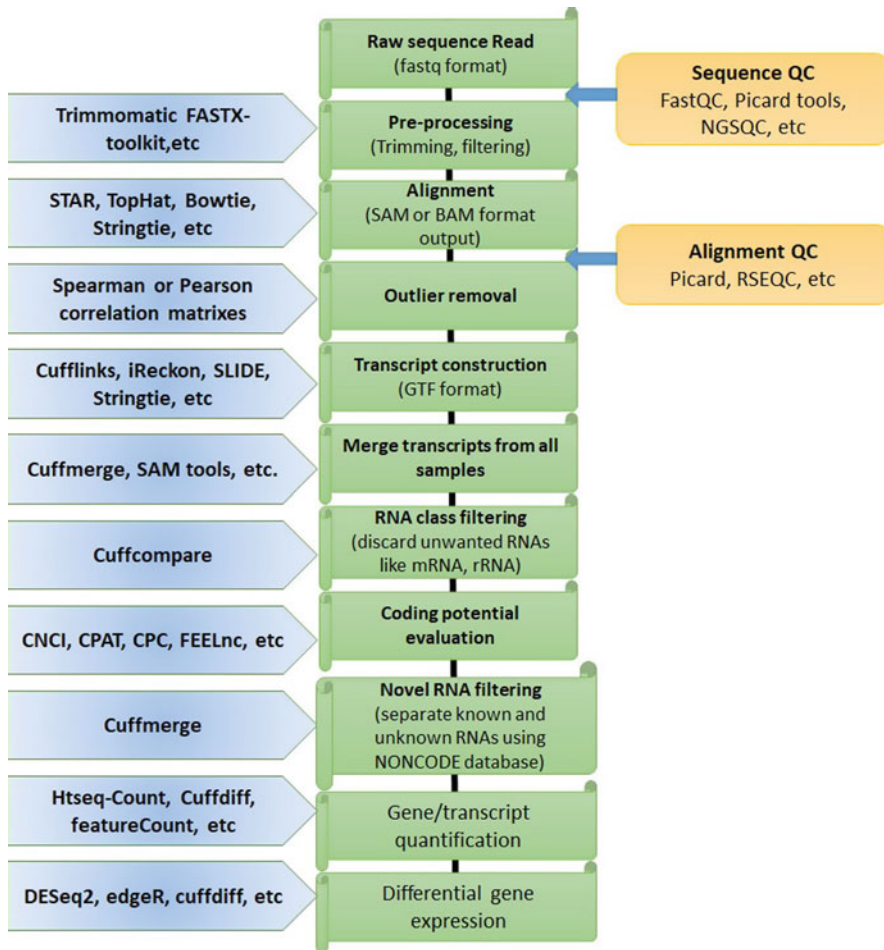


Fig. 12.2 Various steps and tools used in RNA sequencing data analysis. (The figure is adopted from [57])

4. *siRNADB*: It is a database about siRNA (small interfering RNA) that contains information about experimentally verified and predicted siRNAs, sequence, and the literature links [81].
5. *piRNABank*: This database is useful for information on piRNA about their sequence, cluster, and homology searching [82].
6. *snoRNA-LBME-db*: It is a database in which information about snoRNAs and scaRNAs are available with respect to their sequence, expression information, as well as predicted targets including base pairing information [83].
7. *Rfam*: It consists of the data about snRNAs, snoRNAs, and miRNAs and about the sequence families of other structural RNAs [84].

Table 12.4 Web-based tools to analyze RNA sequencing data

Sr no.	Tools	Descriptions	URL	References
1.	DARIO	Used for quantification and annotation of ncRNA	http://dario.bioinf.uni-leipzig.de/	[58]
2.	ShortStack	Used for quantification and annotation of miRNAs	http://axtell-lab-psu.weebly.com/shortstack.html .	[59]
3.	ncPRO-seq	Used for detection of known small ncRNAs and also to discover novel ncRNA species	http://ncpro.curie.fr/	[60]
4.	miRDeep, miRanalyzer	Used for detection of both known and novel microRNAs in small RNA sequencing data	http://59.79.168.90/mirtools http://bioinfo2.ugr.es/miRanalyzer/miRanalyzer.php	[61, 62]
5.	miREval 2.0	Used for detection of novel microRNAs	http://tagc.univ-mrs.fr/mireval	[63]
6.	miRNAkey	Used as a base station for the execution of the first step of analysis of deep sequencing data	http://ibis.tau.ac.il/miRNAkey	[64]
7.	CAP-miRseq	Used for systemic identification of miRNAs	http://bioinformaticstools.mayo.edu/research/cap-mirseq/ .	[65]
8.	tRF2Cancer	Used for the detection of tRNA-derived small RNA fragments (tRFs) and their expression in different cancers	http://rna.sysu.edu.cn/tRFfinder/	[66]
9.	tDRmapper	Used for quantification of tRFs	https://github.com/sararselitsky/tDRmapper	[67]
10.	PhaseTank	Used to detect phasiRNAs	http://phasetank.sourceforge.net/ .	[68]
11.	miRge2.0	Used for differential expression analysis of miRNAs	https://github.com/mhalushka/miRge	[69]

8. *NONCODE*: This database contains data about miRNAs, piRNAs, snoRNAs, and scaRNAs. NONCODE is linked with GenBank [85].
9. *RNAdb*: There is information about miRNAs, piRNAs, snoRNAs, and other ncRNAs that has their sequence with links to literature and other databases[86].
10. *deepBase*: It is a miRNA, piRNA, endo-siRNA, nasRNA, pasRNA, easRNA, and rasRNA database with the data information about their sequence from different tissues and for computationally predicted snRNAs [87].
11. *fRNAdb*: It consists of data about various annotated and predicted ncRNAs of various lengths from different sources [88].

Table 12.5 Various integrated tools to analyze RNA sequencing data

S. no.	Tools	Description	Links	References
1.	CPSS2.0	Used for analyzing small RNA deep sequencing data	https://mcg.ustc.edu.cn/bsc/cps/	[70]
2.	mirTools 2.0	Users to perform noncoding RNA profiling, miRNA target prediction, function annotation of miRNAs targets	http://www.wzgenomics.cn/mr2_dev/news.php	[71]
3.	Oasis 2.0	Used for the analysis of differential expression and classification of small RNAs in deep sequencing data	https://oasis.dzne.de	[72]
4.	The UEA sRNA workbench	Complete analysis of single or multiple-sample small RNA datasets	http://srna-workbench.cmp.uea.ac.uk/	[73]
5.	sRNAtoolbox	High-throughput small RNA profiling and allows to predict novel microRNAs	https://bioinfo5.ugr.es/srnatoolbox	[74]
6.	SPAR	Used for interactive analyses and visualization of small RNA sequencing data	https://spar.lisanwanglab.org/	[75]
7.	sRNAAnalyzer	It is a pipeline for small RNA sequencing data analysis	http://srnanalyzer.systemsbiology.net/	[76]
8.	Unitas	Annotation of small noncoding RNA sequence datasets	https://www.smallmgroup.uni-mainz.de/software.html	[77]

12.7 Applications of Various Transcriptomic Approaches in Context of Dengue and Other Viruses

Expressed sequence tags (ESTs) emerged as the first high-throughput technique to gene expression and annotation of genome. To investigate host gene expression response upon viral infection, EST approach was used as molecular tool. Barón et al. in 2010 compared the differential gene expression profile in the midgut of refractory and susceptible lines of *Aedes aegypti* mosquitoes by infecting with dengue-2 virus [89]. The annotation of EST-identified differentially expressed genes belongs to immune response and metabolism. Another EST-based study by Guo et al. in 2009 in response dengue-2 virus infection in *Aedes albopictus* was analyzed. They identified seven ESTs; among them five were overexpressed in susceptibility, and two were overexpressed in refractory lines [90].

Microarray analysis is regularly used for the analysis of altered gene expression upon virus infections. It is often used in clinical diagnosis to detect the presence of known viruses or to find new viruses [27, 28]. In addition, microarray has been used in the detection and identification of seven pathogenic viruses from the *Flaviviridae*

family; these viruses are the yellow fever (YF), West Nile (WN), Japanese encephalitis (JE), and the dengue strain 1–4 viruses [91]. DNA microarray was used for the detection of dual infection with different DENV serotypes [92]. Likewise, high-density resequencing microarrays (RMAs) are another variant of microarray which has been used for rapid identification and molecular analysis of bacteria and viruses. It also enables to predict for the biomarkers during different clinical outcomes of dengue infection [93, 94]. RMA, microarray technique, has also been proven useful for the rapid and accurate identification of pathogens, specifically for the *Rhabdoviridae* family [95]. However, this technology often has certain limitations in the context of sensitivity, specificity, and reproducibility [38]. Tiling array which is a variant of microarray has a wide range of applications and had been used widely in the detection of multiple foodborne RNA viruses including multiple coxsackievirus serotype (A and B) strains and multiple hepatitis A virus genotype strains. It has also been used in differentiating the virus serotypes [96]. Further, the customized tiling array technique has been utilized into accurate identification of non-polyadenylated RNA in the case of human cytomegalovirus (HCMV) [97].

Transcriptome approach is also used to investigate transcriptional patterns associated with the dengue progression by analyzing the RNA-Seq of peripheral blood mononuclear cells. Patients with varying severity of dengue infection were compared with the patients with other febrile illnesses (OFIs) and the healthy controls. Researchers have collected the sample from individuals; RNA sequencing was performed by constructing a library using Illumina HiSeq 2500. Further, analysis of their data reveals the direct molecular mechanism of bleeding due to decreased platelet count in dengue patients [98]. RNA seq has been used in various studies to explore the potential of noncoding RNA as biomarkers and antiviral targets by analyzing the expression profiles of various noncoding RNAs during different viral infections [99–105]. An RNA seq study was carried out by Castillo et al. [2], where miRNA profiling in DENV-infected primary macrophages was done which identifies miR-3614-5p as an antiviral target [2]. piRNA profiling was done in DENV-infected Asian tiger mosquito and midgut [101]. On the other hand, differential expression of lncRNA during DENV infection was also carried out in which lncRNA was proven to be the potential biomarkers for disease progression [106].

Next-generation sequencing has been widely used nowadays for profiling and discovery of small noncoding RNAs in various infectious diseases. It has been used for profiling of small noncoding RNAs in dengue virus-infected *Aedes* mosquito cells to discover specific viral small noncoding RNAs that facilitate viral replication [107].

Illumina has been widely used to determine an active small interfering RNA-based antiviral response in case of West Nile virus-infected mosquitos [108] and to discover microRNA-like small RNA which autoregulates replication of dengue-2 virus in mosquito cells [109]. A study by Samuel et al. in 2018 carried out multiplex sequencing of DENV serotypes using Illumina and nanopore sequencing simultaneously. They found that multiplex sequencing was robust and generated full genome coverage for all DENV isolates in first attempt unlike single plex approach which failed to produce several amplicons [55]. In a meta-analysis study,

microarray, Affymetrix, and Illumina datasets were compared which revealed the modulation of extracellular matrix and cell junctions during DENV infection which might prove to be the gene signatures of dengue infection [110]. Illumina has been also used to sequence tRNA-derived fragments during infection of respiratory syncytial viral infection [111]. In HIV-infected cells, novel miRNAs and piRNAs have been identified by using next-generation sequencing platform illumine [112].

With the advancements in omics technologies, third-generation sequencing platforms like nanopore can be used for direct sequencing of RNA. Analysis of viral transcriptome using conventional RNA sequencing methods is more complicated because of complexity in splicing patterns, overlapping reading frames, and high gene density. To overcome these problems, Depledge et al. in 2019 used nanopore sequencing to directly sequence RNA from host and viral RNA to study host-pathogen interaction during herpes simplex virus infection [113].

12.8 Conclusion and Future Perspectives

Before 2005, noncoding RNA field was very much scarcely studied. But with the emerging next-generation sequencing technologies, the number of studies increased in the last decade. Various tools and techniques have been introduced to make the ncRNA study to be more comprehensive. This chapter has reviewed the current and emerging approaches for decoding the spectrum of ncRNA functions with the primary focus on virus-related tools and techniques. The field of miRNA is the extensively studied field followed by siRNA and piRNA. While various bioinformatics tools and databases are available for miRNA, siRNA, piRNA, snoRNA, etc., for many ncRNAs, there is a lack of tools and databases for their study which primarily includes lncRNA, circRNA, virus-associated RNAs, etc. This could be one reason for impeding the associated research. Moreover, several tools, techniques, and databases have recently been introduced which help to dig deeper into the possible functions of ncRNAs. Due to extensive expansion of various transcriptome analysis tools, it is expected that various roles of noncoding RNAs are to be discovered in the future which will lead to better understanding of various virus pathophysiology and development of noncoding RNA-based diagnosis and therapeutics.

References

1. Wu S, He L, Li Y, Wang T, Feng L, Jiang L, Zhang P, Huang X (2013) MiR-146a facilitates replication of dengue virus by dampening interferon induction by targeting TRAF6. *J Infect* 67 (4):329–341. <https://doi.org/10.1016/j.jinf.2013.05.003>
2. Castillo JA, Castrillón JC, Dioso-Toro M, Betancur JG, St Laurent G, Smit JM, Urcuqui-Inchima S (2016) Complex interaction between dengue virus replication and expression of miRNA-133a. *BMC Infect Dis* 16(1):29. <https://doi.org/10.1186/s12879-016-1364-y>

3. Firmino N, Martinez VD, Rowbotham DA, Enfield KSS, Bennewith L, Lam WL (2017) HPV status is associated with altered PIWI-interacting RNA expression pattern in head and neck cancer. *Oral Oncol* 55:43–48. <https://doi.org/10.1016/j.oraloncology.2016.01.012>.HPV
4. Léger P, Lara E, Jagla B, Sismeiro O, Mansuroglu Z, Coppée JY, Bonnefoy E, Bouloy M (2013) Dicer-2- and Piwi-mediated RNA interference in Rift Valley fever. *J Virol* 87 (3):1631–1648. <https://doi.org/10.1128/JVI.02795-12>
5. Wang Q, Lee I, Ren J, Ajay SS, Lee YS (2013) Identification and functional characterization of tRNA-derived RNA fragments (tRFs) in respiratory syncytial virus infection. *Mol Ther* 21 (2):368–379. <https://doi.org/10.1038/mt.2012.237>
6. Schorn AJ, Gutbrod MJ, LeBlanc C, Martienssen R (2017) LTR-retrotransposon control by tRNA-derived small RNAs. *Cell* 170(1):61–71.e11. <https://doi.org/10.1016/j.cell.2017.06.013>
7. Watkins NJ, Bohnsack MT (2012) The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA. *WIREs RNA* 3 (June):397–414. <https://doi.org/10.1002/wrna.117>
8. Falaleeva M, Surface J, Shen M, de la Grange P, Stamm S (2015) SNORD116 and SNORD115 change expression of multiple genes and modify each other's activity. *Gene* 572(2):266–273
9. Yu F, Bracken CP, Pillman KA, Lawrence DM, Goodall GJ, Callen DF, Neilsen PM (2015) p53 represses the oncogenic Sno-MiR-28 derived from a SnoRNA. *PLoS One* 10(6):1–20. <https://doi.org/10.1371/journal.pone.0129190>
10. Ma X, Han N, Shao C, Meng Y (2017) Transcriptome-wide discovery of PASRs (promoter-associated small RNAs) and TASRs (terminus-associated small RNAs) in *Arabidopsis thaliana*. *PLoS One* 12(1):1–22. <https://doi.org/10.1371/journal.pone.0169212>
11. Seila AC, Calabrese JM, Levine SS, Yeo GW, Peter B, Flynn RA, Young RA, Sharp PA (2009) Divergent transcription from active promoters. *Science* 322(5909):1849–1851. <https://doi.org/10.1126/science.1162253>.Divergent
12. Wang M, Beck CR, English AC, Meng Q, Buhay C, Han Y, Doddapaneni HV, Yu F, Boerwinkle E, Lupski JR, Muzny DM, Gibbs RA (2015a) PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC Genomics* 16(1):1–12. <https://doi.org/10.1186/s12864-015-1370-2>
13. Wang Y, Yao JIE, Meng H, Yu Z, Wang Z, Yuan X, Chen H, Wang A (2015b) A novel long non-coding RNA, hypoxia-inducible factor-2 α promoter upstream transcript, functions as an inhibitor of osteosarcoma stem cells in vitro. *Mol Med Rep* 11:2534–2540. <https://doi.org/10.3892/mmr.2014.3024>
14. Fan J, Cheng M, Chi X, Liu X, Yang W (2019) A human Long non-coding RNA LncATV promotes virus replication through restricting RIG-I – mediated innate immunity. *Front Immunol* 10(July):1–8. <https://doi.org/10.3389/fimmu.2019.01711>
15. Zhang L, Xu L, Zhang C, Lu Y, Ji T, Ling L (2017) uc38 induces breast cancer cell apoptosis via PBX1. *Am J Cancer Res* 7(12):2438–2451
16. Iiott NE, Heward JA, Roux B, Tsiotou E, Fenwick PS, Lenzi L, Goodhead I, Hertz-Fowler C, Heger A, Hall N, Donnelly LE, Sims D, Lindsay MA (2014) Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes. *Nat Commun* 5:3979. <https://doi.org/10.1038/ncomms4979>
17. Greene J, Baird A, Brady L, Lim M, Gray SG, Mcdermott R, Finn SP (2017) Circular RNAs: biogenesis, function and role in human diseases. *Front Mol Biosci* 4(June):1–11. <https://doi.org/10.3389/fmolb.2017.00038>
18. Tagawa T, Gao S, Koparde VN, Gonzalez M, Spuge JL, Serquiña AP, Lurain K, Ramaswami R, Uldrick TS, Yarchoan R, Ziegelbauer JM (2018) Discovery of Kaposi's sarcoma herpesvirus-encoded circular RNAs and a human antiviral circular RNA. *Proc Natl Acad Sci U S A* 115(50):12805–12810. <https://doi.org/10.1073/pnas.1816183115>
19. Göertz GP, van Bree JWM, Hiralal A, Fernhout BM, Steffens C, Boeren S, Visser TM, Vogels CBF, Abbo SR, Fros JJ, Koenraadt CJM, van Oers MM, Pijlman GP (2019) Subgenomic flavivirus RNA binds the mosquito DEAD/H-box helicase ME31B and determines Zika virus

- transmission by *Aedes aegypti*. Proc Natl Acad Sci U S A 116(38):19136–19144. <https://doi.org/10.1073/pnas.1905617116>
20. Schuessler A, Funk A, Lazear HM, Cooper DA, Torres S, Daffis S, Jha BK, Kumagai Y, Takeuchi O, Hertzog P, Silverman R, Akira S, Barton DJ, Diamond MS, Khromykh AA (2012) West Nile virus noncoding subgenomic RNA contributes to viral evasion of the type I interferon-mediated antiviral response. J Virol 86(10):5708–5718. <https://doi.org/10.1128/jvi.00207-12>
 21. Prusty BK, Gulve N, Govind S, Krueger GRF, Feichtinger J, Larcombe L, Aspinall R, Ablashi DV, Toro CT (2018) Active HHV-6 infection of cerebellar Purkinje cells in mood disorders. Front Microbiol 9(August):1–12. <https://doi.org/10.3389/fmicb.2018.01955>
 22. Takahashi H et al (2014) 5' end-centered expression profiling using cap-analysis gene expression (CAGE) and next-generation sequencing Hazuki. Nature Protocol 7(3):542–561. <https://doi.org/10.1038/nprot.2012.005.5>
 23. Schena M, Shalon D, Davis RW, Brown P (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(October):467–470
 24. Searfoss GH, Ryan TP, Jolly RA (2005) The role of transcriptome analysis in pre-clinical toxicology. Curr Mol Med 5(February 2015):52–64
 25. Mutz K, Heiligenbrinker A, Lo M, Stahl F (2013) Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol 24:22–30. <https://doi.org/10.1016/j.copbio.2012.09.004>
 26. Huyghe A, Francois P, Schrenzel J (2020) Characterization of microbial pathogens by DNA microarray. Infect Genet Evol 9(January):987–995. <https://doi.org/10.1016/j.meegid.2008.10.016>
 27. Parnell GP, Tang BM, Nalos M, Armstrong NJ, Huang SJ, Booth DR, Mclean AS (2013) Identifying key regulatory genes in the whole blood of septic patients to monitor underlying immune dysfunctions. Shock 40(3):166–174. <https://doi.org/10.1097/SHK.0b013e31829ee604>
 28. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, Derisi JL (2002) Microarray-based detection and genotyping of viral pathogens. PNAS 99(24):15687–15692
 29. Wang D, Urisman A, Liu Y, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J, Latreille JP, Wilson RK, Ganem D, Derisi JL (2003) Viral discovery and sequence recovery using DNA microarrays. PLoS Biol 1(2):257–260. <https://doi.org/10.1371/journal.pbio.0000002>
 30. Bertone P, Gerstein M, Snyder M (2005) Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. Chromosom Res 13:259–274
 31. Yazaki J et al (2009) Mapping the genome landscape using tiling array technology. Curr Opin Plant Biol 10(5):534–542. <https://doi.org/10.1016/j.pbi.2007.07.006.Mapping>
 32. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14:331–342. <https://doi.org/10.1101/gr.2094104.Sanger>
 33. Lawrence JG, Hendrix RW, Lawrence JG, Hendrix RW (2001) Where are the pseudogenes in bacterial genomes? Trends Microbiol 9(11):535–540
 34. Buck MJ, Lieb JD (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83(3):349–360. <https://doi.org/10.1016/j.ygeno.2003.11.004>
 35. Zhu H et al (1998) Cellular gene expression altered by human cytomegalovirus: global monitoring with oligonucleotide arrays. Microbiology 95(November):14470–14475
 36. de Saizieu A et al (1998) Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. Nat Biotechnol 16:45–48
 37. Li H (2001) Cluster-Rasch models for microarray gene expression data. Genome Biol 2:1–13

38. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, B. EL. (1996) Expression monitoring by hybridization to high-density oligo nucleotide arrays. *Nat Biotechnol* 14:1675–1680
39. Murphy D (2020) Gene expression studies using microarrays. *Adv Physiol Educ* 26 (4):256–270
40. Choi C, Yoon S, Moon H, Bae Y, Kim C, Diskul-na-ayudthaya P, Van Ngu T, Munir J, Han J, Park S, Moon J, Song S, Ryu S (2018) mirRICH, a simple method to enrich the small RNA fraction from over-dried RNA pellets. *RNA Biol* 6286:763–772. <https://doi.org/10.1080/15476286.2018.1451723>
41. Chaudhary S, Chaudhary PS, Vaishnani TA (2016) Small RNA extraction using fractionation approach and library preparation for NGS platform. *J Adv Res Biotechnol* 1:7
42. Jin G, Cid M, Poole CB, McReynolds LA (2010) Protein mediated miRNA detection and siRNA enrichment using p19. *BioTechniques* 48(6):xvii–xxiii. <https://doi.org/10.2144/000113364>
43. Grentzinger T, Oberlin S, Schott G, Handler D, Svozil J, Barragan-borrero V, Humbert A, Duharcourt S, Brennecke J, Voynet O (2020) A universal method for the rapid isolation of all known classes of functional silencing small RNAs. *Nucleic Acid Res* 48(14):1–15. <https://doi.org/10.1093/nar/gkaa472>
44. Becker C, Hammerle-Fickinger A, Riedmaier I, Pfaffl MW (2010) mRNA and microRNA quality control for RT-qPCR analysis. *Methods* 50(4):237–243. <https://doi.org/10.1016/j.ymeth.2010.01.010>
45. Nilsen TW (2013) Splinted ligation method to detect small RNAs. Cold Spring Harbor Laboratory Press, Harbor, NY, pp 54–59. <https://doi.org/10.1101/pdb.prot072611>
46. CaiFu C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR, Lao KQ, Livak KJ, Guegler KJ (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res* 33(20):1–9. <https://doi.org/10.1093/nar/gni178>
47. Fakruddin M, Chowdhury A (2012) Pyrosequencing-an alternative to traditional sanger sequencing. *Am J Biochem Biotechnol* 8(1):14–20. <https://doi.org/10.3844/ajbbsp.2012.14.20>
48. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* 17(8):1195–1201. <https://doi.org/10.1101/gr.6468307>
49. Merriman B, Torrent I, Rothberg JM (2012) Progress in ion Torrent semiconductor chip based sequencing. *Electrophoresis* 33(23):3397–3417. <https://doi.org/10.1002/elps.201200424>
50. Fujimoto M, Moyerbrailean GA, Noman S, Gizicki JP, Ram ML, Green PA, Ram JL (2014) Application of ion torrent sequencing to the assessment of the effect of alkali ballast water treatment on microbial community diversity. *PLoS One* 9(9):1–9. <https://doi.org/10.1371/journal.pone.0107534>
51. Gamez S, Antoshechkin I, Mendez-Sanchez SC, Akbari OS (2020) The developmental transcriptome of *aedes albopictus*, a major worldwide human disease vector. *G3: Genes, Genomes*. *Genetics* 10(3):1051–1062. <https://doi.org/10.1534/g3.119.401006>
52. Amar L, Benoit C, Beaumont G, Vacher CM, Crepin D, Taouis M, Baroin-Tourancheau A (2012) MicroRNA expression profiling of hypothalamic arcuate and paraventricular nuclei from single rats using Illumina sequencing technology. *J Neurosci Methods* 209(1):134–143. <https://doi.org/10.1016/j.jneumeth.2012.05.033>
53. Ondov BD, Varadarajan A, Passalacqua KD, Bergman NH (2008) Efficient mapping of applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics* 24(23):2776–2777. <https://doi.org/10.1093/bioinformatics/btn512>
54. Li S, Wang H, Qi Y, Tu J, Bai Y, Tian T, Huang N, Wang Y, Xiong F, Lu Z, Xiao Z (2011) Assessment of nanomaterial cytotoxicity with SOLiD sequencing-based microRNA expression profiling. *Biomaterials* 32(34):9021–9030. <https://doi.org/10.1016/j.biomaterials.2011.08.033>

55. Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* 17(1):1–11. <https://doi.org/10.1186/s13059-016-1103-0>
56. Stubbs S, Blacklaws B, Yohan B, Yudhaputri F, Schwem B, Salvaña E, Destura R, Myint K, Sasmono RT, Frost S (2018) A Nanopore-based method for generating complete coding region sequences of dengue virus in resource-limited settings. *Virology*. <https://doi.org/10.1101/499111>
57. Rhoads A, Au KF (2015) PacBio sequencing and its applications. *Genom Proteom Bioinf* 13(5):278–289. <https://doi.org/10.1016/j.gpb.2015.08.002>
58. Ruano P, Delgado LL, Picco S, Villegas L, Tonelli F, Merlo M, Rigau J, Diaz D, Masuelli M (2016) We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists TOP 1%. Intech, tourism, 13. <https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics>
59. Fasold M, Langenberger D, Binder H, Stadler PF, Hoffmann S (2011) DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acid Res* 39(May):112–117. <https://doi.org/10.1093/nar/gkr357>
60. Axtell MJ (2013) ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* 19:740–751. <https://doi.org/10.1261/rna.035279.112.0f>
61. Chen C-j, Servant N, Toedling J, Sarazin A, Marchais A, Duvernois-berthet E, Colot V, Voinnet O, Heard E, Ciaudo C, Barillot E, Strasbourg D, Strasbourg F, Upr C (2012) ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data. *Bioinformatics* 28(23):3147–3149. <https://doi.org/10.1093/bioinformatics/bts587>
62. Aransay AM, Hackenberg M (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acid Res* 39(April):132–138. <https://doi.org/10.1093/nar/gkr247>
63. Williamson V, Kim A, Xie B, McMichael GO, Gao Y, Vladimirov V (2012) Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. *Brief Bioinform* 14(1):36–45. <https://doi.org/10.1093/bib/bbs010>
64. Gao D, Middleton R, Rasko JEJ, Ritchie W (2013) Bioinformatics applications note Sequence analysis miREval 2.0: a web tool for simple microRNA prediction in genome sequences. *Bioinformatics* 29(24):3225–3226. <https://doi.org/10.1093/bioinformatics/btt545>
65. Ronen R, Gan I, Modai S, Sukacheov A, Dror G, Halperin E, Shomron N (2010) miRNAkey: a software for microRNA deep sequencing analysis. *Bioinformatics* 26(20):2615–2616. <https://doi.org/10.1093/bioinformatics/btq493>
66. Sun Z, Evans J, Bhagwate A, Middha S, Bockol M, Yan H, Kocher J (2014) CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. *BMC Genomics* 15:1–10
67. Zheng L, Xu W, Liu S, Sun W, Li J, Wu J, Yang J, Qu L (2016) tRF2Cancer: a web server to detect tRNA-derived small RNA fragments (tRFs) and their expression in multiple cancers. *Nucleic Acid Res* 44(May):185–193. <https://doi.org/10.1093/nar/gkw414>
68. Selitsky SR, Sethupathy P (2015) tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. *BMC Bioinform* 16:1–13. <https://doi.org/10.1186/s12859-015-0800-0>
69. Guo Q, Qu X, Jin W (2015) PhaseTank: genome-wide computational identification of phasiRNAs and their regulatory cascades. *BBA – Gene Regul Mech* 31(2):284–286. <https://doi.org/10.1093/bioinformatics/btu628>
70. Lu Y, Baras AS, Halushka MK (2018) miRge 2.0 for comprehensive analysis of microRNA sequencing data. *BMC Bioinformatics* 19:1–12
71. Wan C, Gao J, Zhang H, Jiang X, Zang Q, Ban R, Zhang Y, Shi Q (2017) Sequence analysis CPSS 2.0: a computational platform update for the analysis of small RNA sequencing data. *Bioinformatics* 33(February):3289–3291. <https://doi.org/10.1093/bioinformatics/btx066>
72. Zhao F, Lang H, Wang Z, Zhang T, Zhang D, Wang R, Lin X, Liu X, Shi P, Pang X (2019) Human novel MicroRNA Seq-915_x4024 in keratinocytes contributes to skin regeneration by

- suppressing scar formation. *Mol Ther Nucleic Acid* 14(March):410–423. <https://doi.org/10.1016/j.omtn.2018.12.016>
73. Rahman R, Gautam A, Bethune J, Sattar A, Fiosins M, Magruder DS, Capece V, Shomroni O, Bonn S (2018) Oasis 2: improved online analysis of small RNA-seq data. *BMC Bioinform* 19:1–10
 74. Stocks MB, Mohorianu I, Beckers M, Paicu C, Moxon S, Thody J, Dalmay T, Moulton V (2018) Sequence analysis the UEA sRNA Workbench (version 4.4): a comprehensive suite of tools for analyzing miRNAs and sRNAs. *Bioinformatics* 34(May):3382–3384. <https://doi.org/10.1093/bioinformatics/bty338>
 75. Cristina G, Rueda A, Barturen G, Lebr R, Oliver L, Hackenberg M (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acid Res* 43(May):467–473. <https://doi.org/10.1093/nar/gkv555>
 76. Valladares O, Kuksa PP, Amlie-wolf A, Wang L, Leung YY (2018) SPAR: small RNA-seq portal for analysis of ~Katani c. *Nucleic Acid Research* 46(May):36–42. <https://doi.org/10.1093/nar/gky330>
 77. Liu Q, Ding C, Lang X, Guo G, Chen J (2019) Small noncoding RNA discovery and profiling with sRNA tools based on high-throughput sequencing. *Brief Bioinform* 00(July):1–11. <https://doi.org/10.1093/bib/bbz151>
 78. Gebert D, Hewel C, Rosenkranz D (2017) Uitas: the universal tool for annotation of small RNAs. *BMC Genomics* 18:1–14. <https://doi.org/10.1186/s12864-017-4031-9>
 79. Bonnal RJP, Rossi RL, Carpi D, Ranzani V, Abrignani S, Pagani M (2015) MiRiadne: A web tool for consistent integration of miRNA nomenclature. *Nucleic Acids Res* 43(W1):W487–W492. <https://doi.org/10.1093/nar/gkv381>
 80. Hausser J, Berninger P, Rodak C, Jantscher Y, Wirth S, Zavolan M (2009) MirZ: an integrated microRNA expression atlas and target prediction resource. *Nucleic Acids Res* 37(Suppl. 2):266–272. <https://doi.org/10.1093/nar/gkp412>
 81. Zhang Y, Zang Q, Xu B, Zheng W, Ban R, Zhang H, Yang Y, Hao Q, Iqbal F, Li A, Shi Q (2016) IsomiR Bank: a research resource for tracking IsomiRs. *Bioinformatics* 32(13):2069–2071. <https://doi.org/10.1093/bioinformatics/btw070>
 82. Truss M, Swat M, Kielbasa SM, Schäfer R, Herzel H, Hagemeyer C (2005) HuSiDa – the human siRNA database: an open-access database for published functional siRNA sequences and technical details of efficient transfer into recipient cells. *Nucleic Acids Res* 33(DATBASE ISS):108–111. <https://doi.org/10.1093/nar/gki131>
 83. Sai Lakshmi S, Agrawal S (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res* 36(Suppl. 1):173–177. <https://doi.org/10.1093/nar/gkm696>
 84. Lestrade L, Weber MJ (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* 34(Database issue):158–162. <https://doi.org/10.1093/nar/gkj002>
 85. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. *Nucleic Acids Res* 31(1):439–441. <https://doi.org/10.1093/nar/gkg006>
 86. Xiyuan L, Dechao B, Liang S, Yang W, Shuangfang F, Hui L, Haitao L, Chunlong L, Wenzheng F, Runsheng C, Yi Z (2017) Using the NONCODE database resource. *Curr Protoc Bioinformatics* 2017:12.16.1–12.16.20. <https://doi.org/10.1002/cpbi.25>
 87. Pang KC, Stephen S, Dinger ME, Engström PG, Lenhard B, Mattick JS (2007) RNAdb 2.0 – an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res* 35(SUPPL. 1):178–182. <https://doi.org/10.1093/nar/gkl926>
 88. Yang JH, Shao P, Zhou H, Chen YQ, Qu LH (2009) deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res* 38(Suppl.1):123–130. <https://doi.org/10.1093/nar/gkp943>
 89. Kin T, Yamada K, Terai G, Okida H, Yoshinari Y, Ono Y, Kojima A, Kimura Y, Komori T, Asai K (2007) iRNAdb: a platform for mining/annotating functional RNA candidates from

- non-coding RNA sequences. *Nucleic Acids Res* 35(SUPPL. 1):145–148. <https://doi.org/10.1093/nar/gkl837>
90. Barón OL, Ursic-bedoya RJ, Lowenberger CA, Ocampo CB (2010) Differential gene expression from midguts of refractory and susceptible lines of the mosquito, *Aedes aegypti*, infected with Dengue-2 virus. *J Insect Sci* 10:1–23
91. Guo X, Yang F, Zhao T, Dong Y, Lu B (2009) Initial analysis of gene expressions in response to dengue-2 virus infection in *Aedes albopictus* (Diptera: Culicidae). *J Agric Urban Entomol* 26(4):157–165. <https://doi.org/10.3954/1523-5475-26.4.157>
92. Falk KI, Lindegren G, Mouzavi-jazi M (2005) DNA microarray technique for detection and identification of seven Flaviviruses pathogenic for man. *J Gen Virol* 540(August):528–540. <https://doi.org/10.1002/jmv.20489>
93. Díaz-Badillo A, Muñoz M d L, Perez-Ramirez G, Altuzar V, Burgueño J, Mendoza-Alvarez JG, Martínez-Muñoz JP, Cisneros A, Navarrete-Espinosa J, Sanchez-Sinencio F (2014) A DNA microarray-based assay to detect dual infection with two dengue virus serotypes. *Sensors (Switzerland)* 14(5):7580–7601. <https://doi.org/10.3390/s140507580>
94. Loke P, Hammond SN, Leung JM, Kim CC, Batra S, Balmaseda A, Harris E (2010) Gene expression patterns of dengue virus-infected children from Nicaragua reveal a distinct signature of increased metabolism. *PLoS Negl Trop Dis* 4(6):1–12. <https://doi.org/10.1371/journal.pntd.0000710>
95. Simmons CP, Popper S, Dolocek C, Nguyen T, Chau B, Griffiths M, Thi N, Dung P, Long TH, Hoang DM, Chau NV, Thi L, Thao T, Hien TT, Relman DA, Farrar J (2007) Patterns of host genome – wide gene transcript abundance in the peripheral blood of patients with acute dengue hemorrhagic fever. *J Infect Dis* 195:1097–1107. <https://doi.org/10.1086/512162>
96. Dacheux L, Berthet N, Dissard G, Holmes EC, Delmas O, Larous F, Guigon G, Dickinson P, Faye O, Sall AA, Old IG, Kong K, Kennedy GC, Manuguerra J, Cole ST, Gessain A (2010) Application of broad-Spectrum resequencing microarray for. *J Virol* 84(18):9557–9574. <https://doi.org/10.1128/JVI.00771-10>
97. Ayodeji M, Kulka M, Jackson SA, Patel I, Mammel M, Cebula TA, Goswami BB (2009) A microarray based approach for the identification of common foodborne viruses. *Open Virol J* 3:7–20
98. Huang Y, Warden C, Zhu H (2013) Accurate identification of non-polyadenylated RNA using a custom human cytomegalovirus tiling array. *J Virol Methods* 187(1):90–93. <https://doi.org/10.1016/j.jviromet.2012.09.020>
99. Banerjee A, Shukla S, Pandey AD, Goswami S, Bandyopadhyay B, Ramachandran V, Das S, Malhotra A, Agarwal A, Adhikari S, Rahman M, Chatterjee S, Bhattacharya N, Basu N, Pandey P, Sood V, Vratil S (2017) RNA-Seq analysis of peripheral blood mononuclear cells reveals unique transcriptional signatures associated with disease progression in dengue patients. *Transl Res* 186:62–78.e9. <https://doi.org/10.1016/j.trsl.2017.06.007>
100. Bao S, Jia L, Zhou X, Zhang ZG, Wu HWL, Yu Z, Ng G, Fan Y, Wong DSM, Huang S, Wang To KK, Yuen KY, Yeung ML, Song YQ (2018) Integrated analysis of mRNA-seq and miRNA-seq for host susceptibilities to influenza A (H7N9) infection in inbred mouse lines. *Funct Integr Genomics*. <https://doi.org/10.1007/s10142-018-0602-3>
101. Barral-Arca R, Gómez-Carballa A, Cebey-López M, Currás-Tuala MJ, Pischedda S, Viz-Lasheras S, Bello X, Martínón-Torres F, Salas A (2020) Rna-seq data-mining allows the discovery of two long non-coding rna biomarkers of viral infection in humans. *Int J Mol Sci*. <https://doi.org/10.3390/ijms21082748>
102. Hess AM, Prasad AN, Ptitsyn A, Ebel GD, Olson KE, Barbacioru C, Monighetti C, Campbell CL (2011) Small RNA profiling of dengue virus-mosquito interactions implicates the PIWI RNA pathway in anti-viral defense. *BMC Microbiol*. <https://doi.org/10.1186/1471-2180-11-45>
103. Lin YT, Kincaid RP, Arasappan D, Dowd SE, Hunicke-Smith SP, Sullivan CS (2010) Small RNA profiling reveals antisense transcription throughout the KSHV genome and novel small RNAs. *RNA*. <https://doi.org/10.1261/ma.1967910>

104. Lu S, Zhu N, Guo W, Wang X, Li K, Yan J, Jiang C, Han S, Xiang H, Wu X, Liu Y, Xiong H, Chen L, Gong Z, Luo F, Hou W (2020) RNA-Seq revealed a circular RNA-microRNA-mRNA regulatory network in Hantaan virus infection. *Front Cell Infect Microbiol*. <https://doi.org/10.3389/fcimb.2020.00097>
105. Shi J, Hu N, Mo L, Zeng Z, Sun J, Hu Y (2018) Deep RNA sequencing reveals a repertoire of human fibroblast circular RNAs associated with cellular responses to herpes simplex virus 1 infection. *Cell Physiol Biochem*. <https://doi.org/10.1159/000491471>
106. Wong AMG, Kong KL, Tsang JWH, Kwong DLW, Guan XY (2012) Profiling of Epstein-Barr virus-encoded microRNAs in nasopharyngeal carcinoma reveals potential biomarkers and oncomirs. *Cancer*. <https://doi.org/10.1002/cncr.26309>
107. Wang XJ, Jiang SC, Wei HX, Deng SQ, He C, Peng HJ (2017) The differential expression and possible function of long noncoding RNAs in liver cells infected by dengue virus. *Am J Trop Med Hyg* 97(6):1904–1912. <https://doi.org/10.4269/ajtmh.17-0307>
108. Miesen P, Ivens A, Buck AH, van Rij RP (2016) Small RNA profiling in dengue virus 2-infected *Aedes Mosquito* cells reveals viral piRNAs and novel host miRNAs. *PLoS Negl Trop Dis* 10(2):1–22. <https://doi.org/10.1371/journal.pntd.0004452>
109. Göertz GP, Fros JJ, Miesen P, Vogels CBF, van der Bent ML, Geertsema C, Koenraadt CJM, van Rij RP, van Oers MM, Pijlman GP (2016) Noncoding subgenomic Flavivirus RNA is processed by the mosquito RNA interference machinery and determines West Nile virus transmission by *Culex pipiens* mosquitoes. *J Virol* 90(22):10145–10159. <https://doi.org/10.1128/jvi.00930-16>
110. Hussain M, Asgari S (2014) MicroRNA-like viral small RNA from dengue virus 2 autoregulates its replication in mosquito cells. *Proc Natl Acad Sci U S A* 111(7):2746–2751. <https://doi.org/10.1073/pnas.1320123111>
111. Afroz S, Giddaluru J, Abbas MM, Khan N (2016) Transcriptome meta-analysis reveals a dysregulation in extra cellular matrix and cell junction associated gene signatures during dengue virus infection. *Sci Rep* 6(September):1–12. <https://doi.org/10.1038/srep33752>
112. Zhou J, Liu S, Chen Y, Fu Y, Silver AJ, Hill MS, Lee I, Lee YS, Bao X (2017) Identification of two novel functional tRNA-derived fragments induced in response to respiratory syncytial virus infection. *J Gen Virol* 98(7):1600–1610. <https://doi.org/10.1099/jgv.0.000852>
113. Chang ST, Thomas MJ, Sova P, Green RR, Palermo RE, Katze MG (2013) Next-generation sequencing of small RNAs from HIV-infected cells identifies phased microRNA expression patterns and candidate novel microRNAs differentially expressed upon infection. *MBio* 4(1):1–10. <https://doi.org/10.1128/mBio.00549-12>
114. Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, Mohr I, Wilson AC (2019) Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat Commun* 10(1):754. <https://doi.org/10.1038/s41467-019-08734-9>



Transcriptomic Approaches in Understanding SARS-CoV-2 Infection

13

Sona Charles and Jeyakumar Natarajan

Abstract

SARS-CoV-2 is a critical disease that has recently acquired the pandemic status worldwide. In order to understand the nature of a pandemic in a short course of time, it is important to elucidate the transcriptomic landscapes of the affected cells and tissues. The high volume of data produced as a result of sequencing-infected samples has a high potential of revealing therapeutic targets or significant pathways that can be clinically exploited for drug or vaccine design. In this chapter we have described the genomic elements and transcriptomic products in SARS-CoV-2 and reviewed a few of the several gene expression studies conducted for determining viral infection mechanisms and response to proposed drugs. We have also outlined pathways of significance in infection and disease progression as well as drug repurposing studies in COVID-19.

Keywords

SARS-CoV-2 · Next-generation sequencing · Transcriptomics · Bioinformatics

13.1 Introduction

The pandemic of COVID-19, first reported to in Wuhan, China, has aroused the concern of human population worldwide as well as the global economy. As of November 9, 2020, a population of 50,743,485 [1] has been affected worldwide. COVID-19 has been declared as a global health emergency as well as a global pandemic by the WHO [2, 3]. On January 30, 2020, the WHO declared the

S. Charles · J. Natarajan (✉)

Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore, Tamilnadu, India

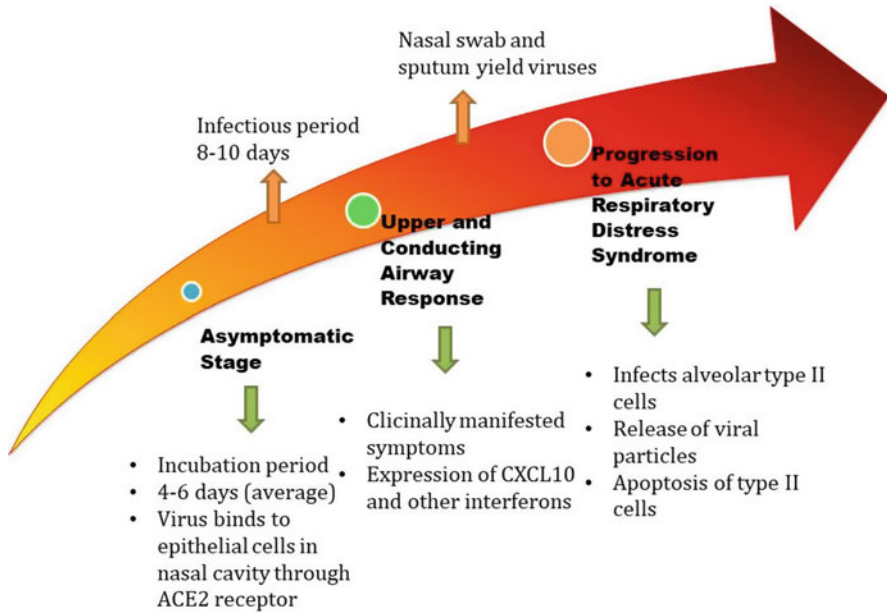


Fig. 13.1 Pathophysiology of SARS-CoV-2

COVID-19 outbreak a global health emergency. On March 11, 2020, the WHO declared COVID-19 a global pandemic, its first such designation since 2009 after declaring H1N1 influenza as a pandemic. COVID-19 is an upper respiratory and gastrointestinal tract infection, with mild or severe symptoms ranging from sore throat, cough and cold, fever, pneumonia and respiratory tract syndrome, and kidney diseases [4]. The primary mode of transmission is via respiratory droplets followed by fomites and aerosols [5]. The average incubation period of the virus is 5.2 days [2, 3]. The epidemic is reported to have started on December 12, 2019, when a novel coronavirus (2019-CoV) that originated in Wuhan and has caused pneumonia-like symptoms [6]. It has led to 1,262,192 deaths (as of November 9, 2020) worldwide [1].

The complexity in pathophysiology (Fig. 13.1) of SARS-CoV-2 has escalated the delay in identifying potential targets and therapeutics against the disease. Since the disease is transmitted through asymptomatic carriers, it is important that the gene/RNA-level differential expression patterns in patients should be well understood. In addition it has been reported in several cases that patients with heart and lung diseases are disproportionately affected which proves that a transcriptome wide analysis is essential to distinguish the clinical manifestations. In order to understand the expression landscapes and signature genes in SARS-CoV-2, it is important to conduct comparative transcriptome-level studies of various cells and tissues in COVID patients and healthy individuals. In this chapter, we will review the genome and transcriptome structure of SARS-CoV-2, computational steps and tools adopted

for gene expression data analysis, recent gene expression studies, and pathways in infection.

13.2 Origin and Transmission

Corona viruses are positive-sense single-stranded RNA viruses. They have the largest genomic structure. They are classified into alpha-, beta-, gamma-, and deltacoronaviruses based on their structure. The origin of alpha- and betacoronaviruses is from bats and rodents, while gamma and delta viruses originated from birds. The novel coronavirus SARS-CoV-2 emerged as a result of recombination events in SARS-related coronaviruses in bats. Civets and humans were infected by the recombined virus and led to the outbreak of SARS-CoV-2 pandemic. Several other strains of coronaviruses also emerged in the past (Table 13.1) such as MERS-CoV, HCoV-229E, HCoV-NL63, HCoV-OC43, HKU1, and swine acute diarrhea syndrome.

13.3 Structure of SARS-CoV-2

SARS-CoV-2 is a single-stranded positive-sense RNA virus [7, 8]. The size of genome is approximately 29.9Kb in size and consists of 13–15 open reading frames [9]. Out of these ORFs, 12 proteins are expressed, whereas 11 encode for proteins. The 5' region of the genome encodes ORF1ab polyproteins, whereas the 3' region codes for S, E, M, and N proteins. Coronaviruses possess four structural proteins, spike (S), membrane (M), envelope (E), and nucleocapsid (N). These proteins play vital roles in assisting virus entry into the host cells and their survival.

SARS-CoV-2 enters the host cell through the endosomal pathway mediated by spike protein, which is a homotrimeric transmembrane glycoprotein [10]. The spike protein consists of two subunits, S1 and S2. S1 binds to the angiotensin-converting enzyme 2 (ACE2), with the help of receptor binding domain, and S2 supports in the fusion of host and viral cell membranes. TMPRSS2 is a cell surface protease that

Table 13.1 Types of coronaviruses and hosts in life cycle

Virus	Natural host	Intermediate	Remarks
Middle East respiratory syndrome coronavirus (MERS-CoV)	Bats	Dromedary camels	Spill over from bats to camels
Human coronavirus 229E (HCoV-229E)	Bats	Camelids	Causes mild infections in immunocompetent humans
Human coronavirus NL63 (HCoV-NL63)	Bats	–	Causes mild infections in immunocompetent humans
HCoV-OC43	Rodents	–	Does not infect humans
HKU1	Rodents		Does not infect humans
HKU2	Bats	Pigs	Does not infect humans

enhances virus entry [11]. The viral RNA is released into the target cell cytoplasm. Upon transcription ORF1a and ORF1ab produces pp1a and pp1ab polyproteins, respectively. The polyproteins undergo cleavage by ORF1a-encoded proteases and produce the components of RNA replicase and transcriptase complex.

E protein which consists of three domains, N terminal hydrophilic domain, hydrophobic transmembrane domain, and a hydrophilic C terminal domain, is the smallest structural protein [12]. The transmembrane region forms pentameric ion channel without ion selectivity and plays the role of a viroporin [13]. It helps in the production of viral particles and their maturation and subsequent release which is influenced by changes in ion homeostasis of components of cells.

The function of N protein is to associate with and package the genetic material and to form self-associated oligomers to form the capsid [14, 15]. These properties are attributed by the N and C terminal domains of the N protein which is intervened by long stretch of disordered region. N protein perturbs the activities of host cell by several mechanisms such as slowing down the cell cycle by inhibiting cyclin E and CDK [16], blocking of cytokinesis by EF1-N protein aggregation inhibiting the function of F-actin [17], inhibiting of interferon production which plays a major role in host immune response [18], inducing of pulmonary fibrosis by TGF β [19], and interacting with host cell proteins. The M protein of SARS-CoV-2 is an integral membrane protein which helps in viral assembly by interacting with the N protein to pack the RNA genome in the capsid [20].

In addition to the structural proteins, ORF3a, ORF6, ORF7a, ORF7b, and ORF8 genes code for accessory proteins [21]. The nonstructural proteins encoded by of SARS-CoV-2 orf1ab and functions of accessory proteins are summarized in Tables 13.2 and 13.3, respectively.

13.4 Host Receptors in Assisting Viral Entry

S protein of SARS-CoV-2 binds to the ACE2 receptor to enter susceptible epithelial cells. The receptors are expressed in a wide array of tissues ranging from the lungs, colon, small intestine, kidney (tubular cells), brain (microglial and neuronal cells), vascular endothelial cells, and cardiac cells [42]. Interaction of ACE2 with SARS-CoV-2 RBD at the interface was determined by X-ray crystallography [43]. TMPRSS is a serine protease that helps in S protein priming [44]. Increased expression of ACE2 and TMPRSS was noted in transcriptomic studies of alveolar epithelial type II cells [45]. The expression pattern of ACE2 and TMPRSS2 were also studied in fetal tissues, and it was not found to be expressed in any tissue except the liver and thymus [45]. Patients with heart failure expressed high levels of ACE2 which explains the severity of infection [46].

Table 13.2 Nonstructural protein in SARS-CoV-2, functions, and host-responsive pathways

Nonstructural protein	Function	Interacting host pathway	Reference
NSP1 (leader protein)	Selective degradation of host mRNA by binding to host cell 40S ribosome	DNA replication Cytoskeleton	[22]
NSP2	NSP2 disrupts host cell by binding to prohibitin 1 and 2 proteins which are involved in cell cycle and differentiation, formation of mitochondria, and cell death	–	[23]
NSP3	Possess papain-like protease domain that cleaves and releases NSP1, NSP2, and NSP3 from orf1a and orf1ab	–	[24]
NSP4	Interaction of NSP4 with NSP3 is necessary for viral replication	Mitochondria	[25]
NSP5	Several mature and intermediate NSPs are generated as a result of cleavage by NSP5	Epigenetic and gene expression regulators	[26]
NSP6	NSP6 prevents degradation of viral particles in the lysosome	Nuclear transport machinery Vesicle trafficking	
NSP7	NSP7 complexes with NSP8 and NSP12 to activate the polymerase action of NSP8	Vesicle trafficking	[27]
NSP8	NSP8-NSP7-NSP12 complex possess RNA polymerase activity	Epigenetic and gene expression regulators RNA processing and regulation Host signaling Mitochondria	[28]
NSP9	NSP9 interacts with DDX5 to assist replication	Nuclear transport machinery Extracellular matrix	[29]
NSP10	Stimulates the activities of NSP14 and NSP16	Vesicle trafficking	[30, 31]
NSP11	Unknown function		
NSP12	RNA-dependent DNA polymerase activity in association with NSP7 and NSP8		[32]
NSP13	Possess helicase activity in the presence of NSP12 and adds 5' cap in viral mRNA	Epigenetic and gene expression regulators Vesicle trafficking Host signaling Cytoskeleton	[33]
NSP14	Possess 3'-5' exoribonuclease activity and N7-methyltransferase activity		[34]
NSP15	Degrades viral polyU sequences and prevents the host immune system from recognizing the virus	Vesicle trafficking Nuclear transport machinery	[35]
NSP16	Prevents recognition by host immune system		[36]

Table 13.3 Non-structural protein in SARS-CoV-2, functions, and host-responsive pathways

Accessory proteins	Function	Interacting host pathway	Reference
Orf3a	Orf3a complexes with TRAF3 and activates ASC ubiquitination and results in caspase 1 activation and IL1 β maturation	–	[37]
Orf6	Orf6 interacts with NSP8 to promote polymerase activity	Nuclear transport machinery	[38]
Orf7a	Type I transmembrane protein	–	[39]
Orf7b	Compartmentalized in Golgi	–	[40]
Orf8b	Orf8 associated with IRF3 and inactivates interferon signaling	Vesicle trafficking	[41]
Orf10	Unknown	Ubiquitin ligases	–

Table 13.4 Bioinformatic tools used for major steps in RNA-Seq data analysis

Function		Tools
Quality checking		FastQC, PRINSEQ
Trimming		Trimmomatic, Cutadapt, Trim Galore
Alignment	Spliced	Bowtie, BWA, Subread, SeqMap, HISAT2, TopHat, STAR, BFAST, MAQ, SOAP2
	Non-spliced	Velvet, Trinity, Mira
Quantification		featureCounts, htseq-count
Normalization and differential expression		limma, edgeR, DESeq2, baySeq, Kallisto, Salmon

13.5 Transcriptomic and Bioinformatic Analysis

The term “transcriptome” refers to the set of all RNA molecules produced in a cell or a tissue at a particular time. The transcriptome is highly dynamic and changes in accordance with the environment, development, as well as during infection. RNA-Seq is a technique which utilizes next-generation sequencing approach to investigate the total RNA content of the cell. The RNA to be sequenced is converted into cDNA fragments and used as a cDNA library. Each fragment is ligated with an adapter sequence and analyzed in NGS. The sequencing may be performed by single-end (sequencing from a single end) or paired-end (sequencing from both ends) methods. Several sequencing platforms are available such as Illumina, Roche 454, and SOLiD which differ in their chemistry and processing. However all these techniques ultimately give reads in the form of FASTQ files. After deriving the FASTQ files, statistical and bioinformatic tools (Table 13.4) are used to extract information from the reads and annotation. The various steps in RNA-Seq data analysis are summarized in Fig. 13.2.

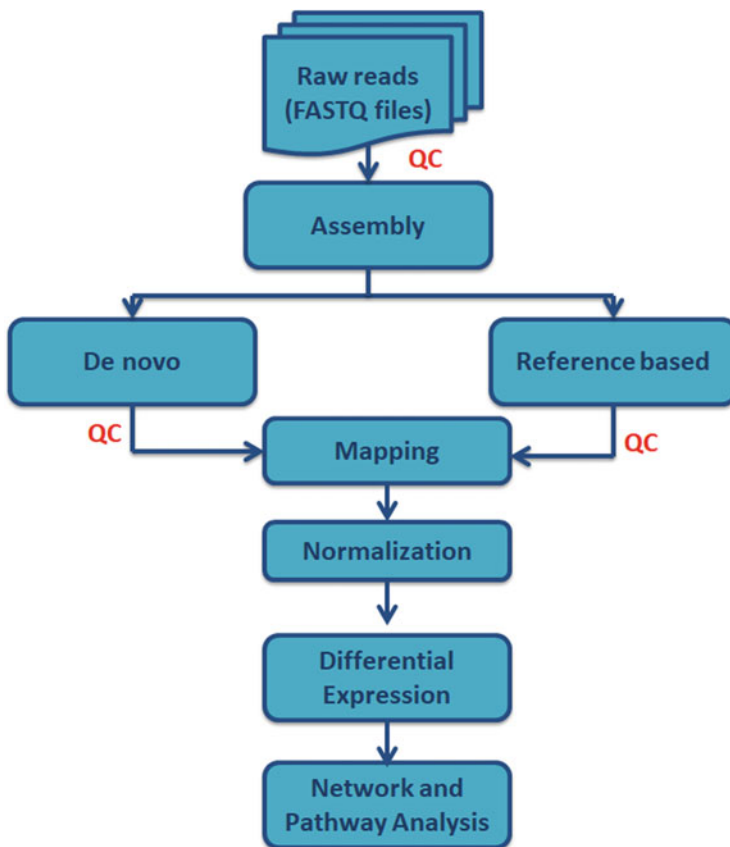


Fig. 13.2 Computational steps in transcriptome data analysis

13.5.1 Preprocessing

The initial step after obtaining the FASTQ files is preprocessing the reads to remove low-quality reads, adapters, and contaminants. The quality of reads is measured by Phred score which can be defined as the probability of inaccuracy in identifying a base at a specific position.

$$Q = -10 \log_{10} P.$$

where Q is the Phred score and P is the probability that a called base is inaccurate.

If the probability of incorrectly called base is 1 in 10, then the base call accuracy will be 90%, and the Phred score will be 10. Higher Phred scores indicate that the quality of called bases is good. Phred scores <30 indicate that the quality of reads is low.

13.5.2 Alignment and Mapping

Alignment is the process of comparing the reads to that of a reference genome. The mapping algorithm matches each read with the reference genome and assigns the read to a particular location, while tolerating certain degree of errors in the alignment. The reads are aligned to a reference genome (if available), and novel unmapped transcripts may be identified. In the absence of reference genome, de novo assembly may be performed by aligning the reads into longer contigs and remapped.

In referenced-based alignment, the reads are aligned with a reference genome or transcriptome. The published reference genomes are deposited in Ensembl, NCBI, and UCSC. The classification of different types of aligners is depicted in Fig. 13.3. Non-spliced aligners are used for prokaryotic genomes since splicing events do not occur. One major problem with eukaryotic genome is splicing which leads to complexity while aligning with reference genome. Several aligners have been developed to overcome this. Spliced aligners use either hash table (dynamic programming) or FM-based (Burrows-Wheeler transform) algorithms.

In de novo assembly, de Bruijn graphs are used to assemble short reads and overlap layout consensus method for long reads [47].

13.5.3 Quantification of Gene Expression

The aligned reads are then used to quantify transcripts or identify transcripts. There are several units for quantifying the transcripts which are given in Table 13.5.

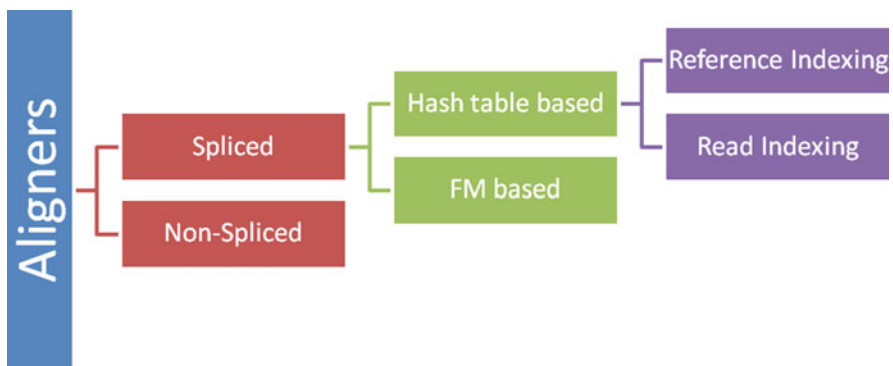


Fig. 13.3 Types of aligners for eukaryotic and prokaryotic reads

Table 13.5 Common units for quantifying gene expression

Units	Meaning
Read count	Number of reads overlapping a particular gene or transcript
Counts per million	$\frac{\text{Number of reads mapped to gene} \times 10^6}{\text{Total number of mapped reads}}$
Reads per kilo base per million mapped reads	$\frac{\text{Number of reads mapped to gene} \times 10^6 \times 10^3}{\text{Number of mapped reads} \times \text{gene length}}$
Transcripts per million	$A \times \frac{1}{\sum(A)} \times 10^6$ where $A = \frac{\text{Total reads mapped to gene} \times 10^3}{\text{genelength}}$

13.5.4 Normalization

In order to make the expression values comparable, normalization methods are used. Though normalization is done based on assumptions, and it is equally important that any errors can have an impact on downstream analysis. The normalized expression values for the reads can be calculated by methods such as total count, upper quartile, median, quantile, trimmed mean of M values, and median of ratios. The impact of normalization methods has been reviewed by several authors, and median of ratio is found to be the best method [48].

13.5.5 Differential Expression Analysis

Differential expression analysis is used to compare the expression of genes or transcripts under specific experimental conditions and to understand the statistical significance of the differences. Several tools have been designed based on statistical models to fit the count data. They have been compared in various studies [49]. The list of tools for differential expression analysis is given in Table 13.5.

However any single method is not suggested but depends on the experimental conditions of the data.

13.6 Transcriptomic Studies in SARS-CoV-2

Infection by SARS-CoV-2 disturbs the host transcriptome and ultimately the proteome. The application of transcriptomics is important in understanding SARS-CoV-2 infection since it enables us to understand the infection process, pathogenesis, as well as drug response. Transcriptomic studies have helped to identify several strategies for host infection by the virus as well as host response to infection. We have briefed a few selected studies as two categories: viral infection mechanisms and drug response.

13.6.1 Viral Infection Mechanisms and Host Response

The first step in understanding the nature of a novel virus is to decode the genome. It is vital to understand the genomic architecture of SARS-CoV-2 to design drugs and vaccines. The genome structure was elucidated by [50]. A detailed analysis of the translational architecture was uncovered with the help of high-throughput sequencing to reveal the mechanisms in a balanced production of viral proteins [51].

Though several studies have thrown light on the impact of COVID-19 on the respiratory system, emerging research is now showing multiple organ dysfunction postinfection. Nearly 10% of COVID-19-affected individuals display gastrointestinal symptoms [52]. ACE2+ cells were found to be present in the digestive tract [53]. Cells having ACE2 receptors are involved in viral entry and are discussed later in the chapter. Increase in inflammatory signaling has been observed in the brain, which suggests that the virus has the potential to cross the blood-brain barrier. The frontal and choroid plexus tissue were shown to contain S protein of virus and antiviral gene IFITM3 [54]. S protein interacted with CD147 to mediate viral invasion, which is a novel route of viral entry [55]. ACE2 and TMPRSS2 have been found in the human olfactory mucosa in a bulk sequencing study which could explain loss of smell in affected patients [56].

A loss of function screen was performed to identify the host factors responsible for infections of alveolar cells by [57]. Their results identified that virus entry can be decreased by sequestration of ACE2 receptors within the cells. A single-cell sequencing study demonstrated the overall immunological landscapes in the host cells, indicating increased acute inflammatory response, activated T cell population, and increased NKTCD56 T cell subset in severe patients. The study indicates a long-term requirement of follow-up changes in the immune system consequent to viral infection. Transcriptional profiling of leukocytes in COVID+ and COVID- patients in the ICU revealed downregulated expression pattern in interferons, genes involved in translation, energetics, blood clotting, complement pathway activation, and TNF pathway [58].

Transcriptomic signatures of infection pattern by respiratory viruses Ebola, H1N1, MERS-CoV, and SARS-CoV were compared to identify the uniqueness of SARS-CoV-2 infection and identified heparin-binding, RAGE, and PLA2 inhibitors, to be associated with SARS-CoV-2 infection [59]. In a recent transcriptomic study, miRNAs were identified with binding sites on SARS-CoV-2 RNA [60]. A meta-analysis study of the transcriptome revealed that hsa-miR-21-3p has a binding site of SARS-CoV-2 RNA and displays increased expression in virus-infected lungs of mouse [61].

13.6.2 Drug Response

Hydroxychloroquine is a controversial drug suggested for use to treat SARS-CoV-2 infections [62]. The mode of action of hydroxychloroquine is not well understood. Rother et al. demonstrated in his study that hydroxychloroquine treatment averts

trained immunity [63]. Trained immunity is the epigenetic reprogramming of cells involved in innate immunity as a result of infection leading to altered response during reinfection. The panacea drug combination hydroxychloroquine-azithromycin is not supported by scientific evidence.

Remdesivir is an adenosine analogue which leads to premature viral transcription termination [64]. However the toxicity of the drug has been reported through the inhibition of mitochondrial RNA polymerase [65]. Expression profiling study by high-throughput sequencing was conducted to identify genes whose loss of activity decreases the toxicity of remdesivir, and it was found that *SLC29A3* compensates the toxicity without losing the antiviral potency of the drug [66].

13.7 Drug Repurposing

Drug repurposing or drug repositioning is an approach to explore the applicability of approved drugs in diseases other than the original scope of medical indication [67]. This reduces the time taken for preclinical studies as the drug has already passed the trials. Data-driven computational approaches are widely used for drug repurposing as depicted in Fig. 13.4.

13.7.1 Computational Docking Studies

Docking studies are based on the complementarity in binding regions of ligand and a target. Virtual screening-based studies utilize multiple compounds from existing databases to interrogate against a target protein. A computation and molecular docking study has been conducted to four major targets, NSP4, RdRp, NendoU/nsp5, and ACE2. They identified Baicalin and Limonin as two leads to target NSP4, RdRp, NendoU/nsp5, and ACE2, respectively [68]. Another in silico study identified that GR 127935 hydrochloride hydrate, GNF-5, RS504393, TNP, and eptifibatid acetate were found to target the ACE2 receptor binding motifs [69]. ACE2 is a promising candidate for docking studies as it is the first contact during entry into the host cells.

13.7.2 Signature-Based Studies

Molecular signatures are a unique characteristic feature of a drug which attributes to its properties. The signature of a drug may be chemical, transcriptomic, or proteomic. Transcriptomic signatures may be compared and matched to identify association of drugs with diseases and other drugs. The drug-disease association experiment may be conducted to identify how the expression of genes is altered with and without a drug when compared to healthy and diseased patient samples. Drug-drug association experiments are based on the principle of guilt by association. Drugs having similar signatures indicate a similar therapeutic impact regardless of

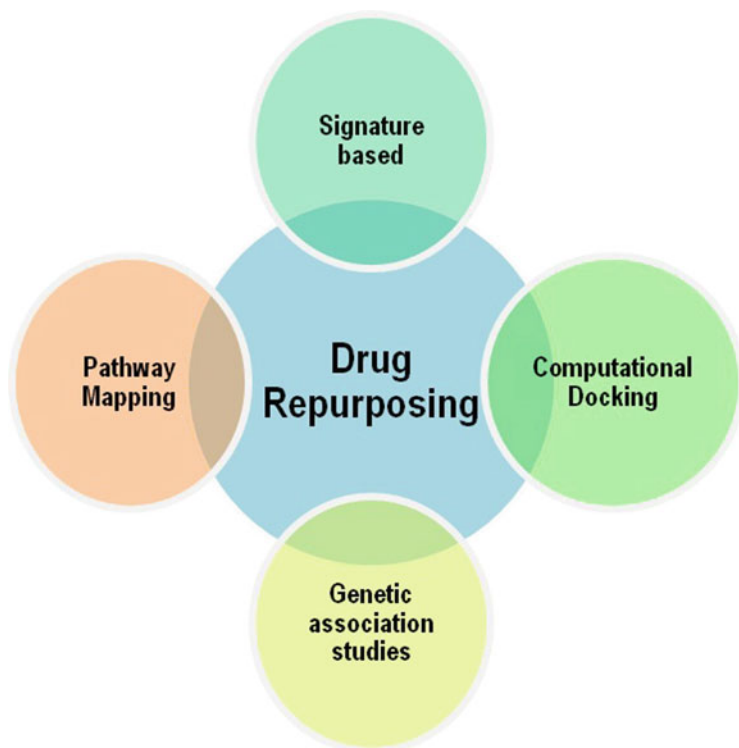


Fig. 13.4 Drug repurposing approaches used in SARS-CoV-2 drug discovery

their structure. Transcriptome-based drug repositioning methods were applied to shortlist prospective candidate drug targets using computational approaches [70].

13.7.3 Genetic Association Studies

Genome-wide association and multi-omic studies are widely used for probing the genetic landscape of diseases. GWAS studies on COVID patients with severe respiratory syndrome have been published. The gene cluster 3p21.31 was found to be potentially associated with respiratory failure. The study also indicated a clear association between disease severity and blood groups, A group being the most affected and O group least affected [7, 8].

13.7.4 Pathway Mapping

Due to the vast inflow of gene expression data into public databases, research is focused on identifying pathways and genes triggered or suppressed after viral

infections. Pathway mapping is the assignment of genes to biological pathways using network-based approaches. Pathway enrichment analysis is a common technique to enrich genes in a specific pathway. These approaches have contributed to the identification of novel genes in a pathway through methods like association rule mining, latent pathway analysis, etc. COVID-19 disease map is an open science collaborative approach to build and update a repository for COVID-19 interactions and pathways [71].

13.8 Infection Pathways Identified from Transcriptomics

As a consequence of viral infection, several host response pathways are activated during different stages of pathogenicity. Infection triggers stress and in response p38MAPK pathway is activated and causes apoptosis. Angiotensin II is converted into angiotensin 1–7 by ACE2. Angiotensin II acts through p38MAPK signaling pathway. Viral entry leads to decrease in the ACE2 activity causing downregulation of angiotensin 1–7 [72]. Therefore angiotensin II results in activation of p38 in the lungs and heart which causes unregulated inflammation. P38 pathway also upregulates the protease ADAM17 which results in cleavage of ACE2 ectodomain, thereby further decreasing local ACE2 activity [73]. The above studies indicate that p38 pathway is a promising target for drug development.

Angiotensin II induces IL6 which is upregulated in COVID-19 patients and increases JAK/STAT pathway leading to inflammation and cytokinin storm. Angiotensin II induces IL6 secretion. S protein downregulates ACE2 causing upregulation of angiotensin II leading to IL6 production through JAK/STAT pathway [74]. The ultimate effect is lung damage due to inflammation. In another transcriptomic study, it was identified that in order to block the production of interferon-stimulated genes, SARS-CoV-2 uses the strategy to block STAT1 and STAT2 transcription factors from being translocated into the nucleus [75].

Pathway analysis using transcriptomic data submitted on public databases through in silico methods has revealed several pathways which require experimental validation for confirmation.

13.9 Conclusion

Gene expression data has a wide potential for being explored for additional data far beyond the objectives for which the experiment is conducted. COVID-19 is a novel disease, and hence knowledge gaps persist in understanding the mechanism and interactions with the host cells. Transcriptomics helps in the detection of novel genes associated with infection mechanism, interaction between host and virus, and possibly personalized treatment in patients categorized into various risk levels. Transcriptomic methods will assist in providing a complete overview of disease progression as well as to identify the *Achilles' heel* in therapeutic intervention. Comparative studies of gene expression profiles will help in staging and also in

suggesting treatment regimes. It can be concluded that the interplay between host and viral mechanisms may be better understood through transcriptomic analysis and bioinformatics.

References

1. Worldometer (2020) Coronavirus cases. Worldometer
2. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M et al (2020a) Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N Engl J Med*. <https://doi.org/10.1056/nejmoa2001316>
3. Li X, Wang W, Zhao X, Zai J, Zhao Q, Li Y, Chaillon A (2020b) Transmission dynamics and evolutionary history of 2019-nCoV. *J Med Virol*. <https://doi.org/10.1002/jmv.25701>
4. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan. *China The Lancet*. [https://doi.org/10.1016/S0140-6736\(20\)30183-5](https://doi.org/10.1016/S0140-6736(20)30183-5)
5. MacKenzie JS, Smith DW (2020) COVID-19: a novel zoonotic disease caused by a coronavirus from China: what we know and what we don't. *Microbiol Australia*. <https://doi.org/10.1071/MA20013>
6. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Di Jiang R, Liu MQ, Chen Y, Shen XR, Wang X et al (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. <https://doi.org/10.1038/s41586-020-2012-7>
7. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020a Mar; 579 (7798):265–269. <https://doi.org/10.1038/s41586-020-2008-3>. Epub 2020 Feb 3. Erratum in: *Nature*. 2020 Apr; 580(7803):E7. PMID: 32015508; PMCID: PMC7094943
8. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020b) A new coronavirus associated with human respiratory disease in China. *Nature*. <https://doi.org/10.1038/s41586-020-2008-3>
9. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L et al (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8)
10. Walls AC, Tortorici MA, Bosch BJ, Frenz B, Rottier PJM, DiMaio F, Rey FA, Veerles D (2016) Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature*. <https://doi.org/10.1038/nature16988>
11. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu NH, Nitsche A, Müller MA, Drosten C, Pöhlmann S (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*. <https://doi.org/10.1016/j.cell.2020.02.052>
12. Surya W, Li Y, Torres J (2018) Structural model of the SARS coronavirus E channel in LMPG micelles. *Biochim Biophys Acta Biomembr*. <https://doi.org/10.1016/j.bbmem.2018.02.017>
13. Ye Y, Hogue BG (2007) Role of the coronavirus E Viroporin protein Transmembrane domain in virus assembly. *J Virol*. <https://doi.org/10.1128/jvi.01472-06>
14. Chen CY, Chang CK, Chang YW, Sue SC, Bai HI, Riang L, Hsiao CD, Huang TH (2007) Structure of the SARS coronavirus Nucleocapsid protein RNA-binding dimerization domain

- suggests a mechanism for helical packaging of viral RNA. *J Mol Biol.* <https://doi.org/10.1016/j.jmb.2007.02.069>
15. Luo H, Chen J, Chen K, Shen X, Jiang H (2006) Carboxyl terminus of severe acute respiratory syndrome coronavirus nucleocapsid protein: self-association analysis and nucleic acid binding characterization. *Biochemistry.* <https://doi.org/10.1021/bi0609319>
 16. Surjit M, Liu B, Chow VTK, Lal SK (2006) The nucleocapsid protein of severe acute respiratory syndrome-coronavirus inhibits the activity of cyclin-cyclin-dependent kinase complex and blocks S phase progression in mammalian cells. *J Biol Chem.* <https://doi.org/10.1074/jbc.M509233200>
 17. Surjit M, Liu B, Jameel S, Chow VTK, Lal SK (2004) The SARS coronavirus nucleocapsid protein induces actin reorganization and apoptosis in COS-1 cells in the absence of growth factors. *Biochem J.* <https://doi.org/10.1042/BJ20040984>
 18. Spiegel M, Pichlmair A, Martínez-Sobrido L, Cros J, García-Sastre A, Haller O, Weber F (2005) Inhibition of Beta interferon induction by severe acute respiratory syndrome coronavirus suggests a two-step model for activation of interferon regulatory factor 3. *J Virol.* <https://doi.org/10.1128/jvi.79.4.2079-2086.2005>
 19. Zhao X, Nicholls JM, Chen YG (2008) Severe acute respiratory syndrome-associated coronavirus nucleocapsid protein interacts with Smad3 and modulates transforming growth factor- β signaling. *J Biol Chem.* <https://doi.org/10.1074/jbc.M708033200>
 20. Siu YL, Teoh KT, Lo J, Chan CM, Kien F, Escriou N, Tsao SW, Nicholls JM, Altmeyer R, Peiris JSM, Bruzzone R, Nal B (2008) The M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles. *J Virol.* <https://doi.org/10.1128/jvi.01052-08>
 21. Khailany RA, Safdar M, Ozaslan M (2020) Genomic characterization of a novel SARS-CoV-2. *Gene Rep.* <https://doi.org/10.1016/j.genrep.2020.100682>
 22. Huang C, Lokugamage KG, Rozovics JM, Narayanan K, Semler BL, Makino S (2011) SARS coronavirus nsp1 protein induces template-dependent endonucleolytic cleavage of mRNAs: viral mRNAs are resistant to nsp1-induced RNA cleavage. *PLoS Pathog.* <https://doi.org/10.1371/journal.ppat.1002433>
 23. Cornillez-Ty CT, Liao L, Yates JR, Kuhn P, Buchmeier MJ (2009) Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. *J Virol.* <https://doi.org/10.1128/jvi.00842-09>
 24. Lei J, Kusov Y, Hilgenfeld R (2018) Nsp3 of coronaviruses: structures and functions of a large multi-domain protein. In *Antiviral Res.* <https://doi.org/10.1016/j.antiviral.2017.11.001>
 25. Sakai Y, Kawachi K, Terada Y, Omori H, Matsuura Y, Kamitani W (2017) Two-amino acids change in the nsp4 of SARS coronavirus abolishes viral replication. *Virology.* <https://doi.org/10.1016/j.virol.2017.07.019>
 26. Tomar S, Johnston ML, John SES, Osswald HL, Nyalapatla PR, Paul LN, Ghosh AK, Denison MR, Mesecar AD (2015) Ligand-induced dimerization of Middle East respiratory syndrome (MERS) coronavirus nsp5 protease (3CLpro): implications for nsp5 regulation and the development of antivirals. *J Biol Chem.* <https://doi.org/10.1074/jbc.M115.651463>
 27. TeVelthuis AJW, Van Den Worm SHE, Snijder EJ (2012) The SARS-coronavirus nsp7+nsp8 complex is a unique multimeric RNA polymerase capable of both de novo initiation and primer extension. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkr893>
 28. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, Wang T, Sun Q, Ming Z, Zhang L, Ge J, Zheng L, Zhang Y, Wang H, Zhu Y, Zhu C, Hu T, Hua T, Zhang B et al (2020) Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science.* <https://doi.org/10.1126/science.abb7498>
 29. Zhao S, Ge X, Wang X, Liu A, Guo X, Zhou L, Yu K, Yang H (2015) The DEAD-box RNA helicase 5 positively regulates the replication of porcine reproductive and respiratory syndrome virus by interacting with viral Nsp9 in vitro. *Virus Res.* <https://doi.org/10.1016/j.virusres.2014.10.021>

30. Ma Y, Wu L, Shaw N, Gao Y, Wang J, Sun Y, Lou Z, Yan L, Zhang R, Rao Z (2015) Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. *Proc Natl Acad Sci U S A*. <https://doi.org/10.1073/pnas.1508686112>
31. Wang Y, Sun Y, Wu A, Xu S, Pan R, Zeng C, Jin X, Ge X, Shi Z, Ahola T, Chen Y, Guo D (2015) Coronavirus nsp10/nsp16Methyltransferase can be targeted by nsp10-derived peptide in vitro and in vivo to reduce replication and pathogenesis. *J Virol*. <https://doi.org/10.1128/jvi.00948-15>
32. Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, Snijder EJ, Canard B, Imbert I (2014) One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. *Proc Natl Acad Sci U S A*. <https://doi.org/10.1073/pnas.1323705111>
33. Ivanov KA, Thiel V, Dobbe JC, van der Meer Y, Snijder EJ, Ziebuhr J (2004) Multiple enzymatic activities associated with severe acute respiratory syndrome coronavirus helicase. *J Virol*. <https://doi.org/10.1128/jvi.78.11.5619-5632.2004>
34. Case JB, Ashbrook AW, Dermody TS, Denison MR (2016) Mutagenesis of S -Adenosyl-l-methionine-binding residues in coronavirus nsp14N7-Methyltransferase demonstrates differing requirements for genome translation and resistance to innate immunity. *J Virol*. <https://doi.org/10.1128/jvi.00542-16>
35. Hackbart M, Deng X, Baker SC (2020) Coronavirus endoribonuclease targets viral polyuridine sequences to evade activating host sensors. *Proc Natl Acad Sci U S A*. <https://doi.org/10.1073/pnas.1921485117>
36. Decroly E, Debarnot C, Ferron F, Bouvet M, Coutard B, Imbert I, Gluais L, Papageorgiou N, Sharff A, Bricogne G, Ortiz-Lombardia M, Lescar J, Canard B (2011) Crystal structure and functional analysis of the SARS-coronavirus RNA cap 2'-o-methyltransferasensp10/nsp16 complex. *PLoS Pathog*. <https://doi.org/10.1371/journal.ppat.1002059>
37. Siu KL, Yuen KS, Castano-Rodriguez C, Ye ZW, Yeung ML, Fung SY, Yuan S, Chan CP, Yuen KY, Enjuanes L, Jin DY (2019) Severe acute respiratory syndrome coronavirus ORF3a protein activates the NLRP3inflammasome by promoting TRAF3-dependent ubiquitination of ASC. *FASEB J*. <https://doi.org/10.1096/fj.201802418R>
38. Kumar P, Gunalan V, Liu B, Chow VTK, Druce J, Birch C, Catton M, Fielding BC, Tan YJ, Lal SK (2007) The nonstructural protein 8 (nsp8) of the SARS coronavirus interacts with its ORF6 accessory protein. *Virology*. <https://doi.org/10.1016/j.virol.2007.04.029>
39. Nelson CA, Pekosz A, Lee CA, Diamond MS, Fremont DH (2005) Structure and intracellular targeting of the SARS-coronavirus orf7a accessory protein. *Structure*. <https://doi.org/10.1016/j.str.2004.10.010>
40. Schaecher SR, Mackenzie JM, Pekosz A (2007) The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into SARS-CoV particles. *J Virol*. <https://doi.org/10.1128/jvi.01691-06>
41. Wong HH, Fung TS, Fang S, Huang M, Le MT, Liu DX (2018) Accessory proteins 8b and 8ab of severe acute respiratory syndrome coronavirus suppress the interferon signaling pathway by mediating ubiquitin-dependent rapid degradation of interferon regulatory factor 3. *Virology*. <https://doi.org/10.1016/j.virol.2017.12.028>
42. Fätsigndriks L (2010) The angiotensin II type 2 receptor and the gastrointestinal tract. *JRAAS – J Renin Angiotensin Aldosterone Syst*. <https://doi.org/10.1177/1470320309347788>
43. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, Zhang Q, Shi X, Wang Q, Zhang L, Wang X (2020) Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*. <https://doi.org/10.1038/s41586-020-2180-5>
44. Glowacka I, Bertram S, Muller MA, Allen P, Soilleux E, Pfefferle S, Steffen I, Tsegaye TS, He Y, Gnirss K, Niemeyer D, Schneider H, Drosten C, Pohlmann S (2011) Evidence that TMPRSS2 activates the severe acute respiratory syndrome coronavirus spike protein for membrane fusion and reduces viral control by the Humoral immune response. *J Virol*. <https://doi.org/10.1128/jvi.02232-10>

45. Zhao Y, Zhao Z, Wang Y, Zhou Y, Ma Y, Zuo W (2020) Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCoVdoi: bioRxiv preprint. BioRxiv
46. Chen L, Li X, Chen M, Feng Y, Xiong C (2020) The ACE2 expression in human heart indicates new potential mechanism of heart injury among patients infected with SARS-CoV-2. *Cardiovasc Res*. <https://doi.org/10.1093/cvr/cvaa078>
47. Compeau PE, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* 29(11):987–991. <https://doi.org/10.1038/nbt.2023>
48. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaeffer B, Le Crom S, Guedj M, Jaffrézic F, French StatOmiQue Consortium (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 14(6):671–683. <https://doi.org/10.1093/bib/bbs046>
49. Sonesson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform* 14:91. <https://doi.org/10.1186/1471-2105-14-91>
50. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H (2020) The architecture of SARS-CoV-2 Transcriptome. *Cell*. <https://doi.org/10.1016/j.cell.2020.04.011>
51. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, Tamir H, Achdout H, Stein D, Israeli O, Beth-Din A, Melamed S, Weiss S, Israely T, Paran N, Schwartz M, Stern-Ginossar N (2020) The coding capacity of SARS-CoV-2. *Nature*. <https://doi.org/10.1038/s41586-020-2739-1>. Advance online publication
52. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui D, Du B, Li LJ, Zeng G, Yuen KY, Chen RC, Tang CL, Wang T, Chen PY, Xiang J, Li SY et al (2020) Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med* 382(18):1708–1720. <https://doi.org/10.1056/NEJMoa2002032>
53. Xu J, Chu M, Zhong F, Tan X, Tang G, Mai J, Lai N, Guan C, Liang Y, Liao G (2020) Digestive symptoms of COVID-19 and expression of ACE2 in digestive tract organs. *Cell Death Discovery* 6(1):76. <https://doi.org/10.1038/s41420-020-00307-w>
54. Yang AC, Kern F, Losada PM, Maat CA, Schmartz G, Fehlmann T, Schaum N, Lee DP, Calcuttawala K, Vest RT, Gate D, Berdnik D, McNerney MW, Channappa D, Cobos I, Ludwig N, Schulz-Schaeffer WJ, Keller A, Wyss-Coray T (2020) Broad transcriptional dysregulation of brain and choroid plexus cell types with COVID-19. *BioRxiv* 2020 (10):22.349415. <https://doi.org/10.1101/2020.10.22.349415>
55. Wang K, Chen W, Zhou Y-S, Lian J-Q, Zhang Z, Du P, Gong L, Zhang Y, Cui H-Y, Geng J-J, Wang B, Sun X-X, Wang C-F, Yang X, Lin P, Deng Y-Q, Wei D, Yang X-M, Zhu Y-M et al (2020) SARS-CoV-2 invades host cells via a novel route: CD147-spike protein. *BioRxiv* 2020 (03):14.988345. <https://doi.org/10.1101/2020.03.14.988345>
56. Brann DH, Tsukahara T, Weinreb C, Lipovsek M, Van den Berge K, Gong B, Chance R, Macaulay IC, Chou HJ, Fletcher RB, Das D, Street K, de Bezieux HR, Choi YG, Risso D, Dudoit S, Purdom E, Mill J, Hachem RA, Matsunami H et al (2020) Non-neuronal expression of SARS-CoV-2 entry genes in the olfactory system suggests mechanisms underlying COVID-19-associated anosmia. *Sci Adv* 6(31):eabc5801. <https://doi.org/10.1126/sciadv.abc5801>
57. Daniloski Z, Jordan TX, Wessels HH, Hoagland DA, Kasela S, Legut M, Maniatis S, Mimitou EP, Lu L, Geller E, Danziger O, Rosenberg BR, Phatnani H, Smibert P, Lappalainen T, tenOever BR, Sanjana NE (2020) Identification of required host factors for SARS-CoV-2 infection in human cells. *Cell*. <https://doi.org/10.1016/j.cell.2020.10.030>. Advance online publication
58. Fraser DD, Cepinskas G, Slessarev M, Martin C, Daley M, Miller MR, O’Gorman DB, Gill SE, Patterson EK, Dos Santos CC (2020) Inflammation profiling of critically ill coronavirus disease 2019 patients. *Crit Care Exp* 2(6):e0144. <https://doi.org/10.1097/CCE.0000000000000144>
59. Alsamman AM, Zayed H (2020) The transcriptomic profiling of COVID-19 compared to SARS, MERS, Ebola, and H1N1. *BioRxiv* 2020(05):06.080960. <https://doi.org/10.1101/2020.05.06.080960>
60. Fulzele S, Sahay B, Yusufu I, Lee TJ, Sharma A, Kolhe R, Isales CM (2020) COVID-19 virulence in aged patients might be impacted by the host cellular MicroRNAs abundance/profile. *Aging Dis*. <https://doi.org/10.14336/AD.2020.0428>

61. Nersisyan S, Engibaryan N, Gorbonos A, Kirdey K, Makhonin A, Tonevitsky A (2020) Potential role of cellular miRNAs in coronavirus-host interplay. *Peer J*. <https://doi.org/10.7717/peerj.9994>
62. Yao X, Ye F, Zhang M, Cui C, Huang B, Niu P, Liu X, Zhao L, Dong E, Song C, Zhan S, Lu R, Li H, Tan W, Liu D (2020) In vitro antiviral activity and projection of optimized dosing design of hydroxychloroquine for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* 71(15):732–739. <https://doi.org/10.1093/cid/ciaa237>
63. Rother N, Yanginlar C, Lindeboom RGH, Bekkering S, van Leent MMT, Buijsers B, Jonkman I, de Graaf M, Baltissen M, Lamers L, Riksen N, Fayad Z, Mulder WJM, Hilbrands L, Joosten LAB, Netea M, Vermeulen M, van der Vlag J, Duijvenvoorden R (2020) Hydroxychloroquine inhibits trained immunity – implications for COVID-19. medRxiv. <https://doi.org/10.1101/2020.06.08.20122143>
64. Warren TK, Jordan R, Lo MK, Ray AS, Mackman RL, Soloveva V, Siegel D, Perron M, Bannister R, Hui HC, Larson N, Strickley R, Wells J, Stuthman KS, Van Tongeren SA, Garza NL, Donnelly G, Shurtleff AC, Retterer CJ et al (2016) Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. *Nature*. <https://doi.org/10.1038/nature17180>
65. Tchesnokov EP, Feng JY, Porter DP, Götte M (2019) Mechanism of inhibition of ebola virus RNA-dependent RNA polymerase by remdesivir. *Viruses*. <https://doi.org/10.3390/v11040326>
66. Akinci E, Cha M, Lin L, Yeo G, Hamilton MC, Donahue CJ, Bermudez-Cabrera HC, Zanetti LC, Chen M, Barkal SA, Khowpinitchai B, Chu N, Velimirovic M, Jodhani R, Fife JD, Sovrovic M, Cole PA, Davey RA, Cassa CA, Sherwood RI (2020) Elucidation of remdesivir cytotoxicity pathways through genome-wide CRISPR-Cas9 screening and transcriptomics. *BioRxiv: The Preprint Server for Biology*. <https://doi.org/10.1101/2020.08.27.270819>
67. Ashburn TT, Thor KB (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 3(8):673–683. <https://doi.org/10.1038/nrd1468>
68. Alazmi M, Motwalli O (2020) In silico virtual screening, characterization, docking and molecular dynamics studies of crucial SARS-CoV-2 proteins. *J Biomol Struct Dyn*, 1–11. Advance online publication. <https://doi.org/10.1080/07391102.2020.1803965>
69. Choudhary S, Malik YS, Tomar S (2020) Identification of SARS-CoV-2 cell entry inhibitors by drug repurposing using in silico structure-based virtual screening approach. *Front Immunol* 11:1664. <https://www.frontiersin.org/article/10.3389/fimmu.2020.01664>
70. Jia Z, Song X, Shi J, Wang W, He K (2020) Transcriptome-based drug repositioning for coronavirus disease 2019 (COVID-19). *Pathog Dis* 78(4):ftaa036. <https://doi.org/10.1093/femspd/ftaa036>
71. Ostaszewski M, Mazein A, Gillespie ME, Kuperstein I, Niarakis A, Hermjakob H, Pico AR, Willighagen EL, Evelo CT, Hasenauer J, Schreiber F, Dräger A, Demir E, Wolkenhauer O, Furlong LI, Barillot E, Dopazo J, Orta-Resendiz A, Messina F et al (2020) COVID-19 disease map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Sci Data* 7(1):136. <https://doi.org/10.1038/s41597-020-0477-8>
72. Park JK, Fischer R, Dechend R, Shagdarsuren E, Gapeljuk A, Wellner M, Meiners S, Gratze P, Al-Saadi N, Feldt S, Fiebeler A, Madwed JB, Schirdewan A, Haller H, Luft FC, Müller DN (2007) p38 mitogen-activated protein kinase inhibition ameliorates angiotensin II-induced target organ damage. *Hypertension (Dallas, Tex 1979)* 49(3):481–489. <https://doi.org/10.1161/01.HYP.0000256831.33459.ea>
73. Scott AJ, O’Dea KP, O’Callaghan D, Williams L, Dokpesi JO, Tatton L, Handy JM, Hogg PJ, Takata M (2011) Reactive oxygen species and p38 mitogen-activated protein kinase mediate tumor necrosis factor α -converting enzyme (TACE/ADAM-17) activation in primary human monocytes. *J Biol Chem*. <https://doi.org/10.1074/jbc.M111.277434>
74. Glowacka I, Bertram S, Herzog P, Pfefferle S, Steffen I, Muench MO, Simmons G, Hofmann H, Kuri T, Weber F, Eichler J, Drosten C, Pöhlmann S (2010) Differential Downregulation of ACE2 by the spike proteins of severe acute respiratory syndrome coronavirus and human coronavirus NL63. *J Virol*. <https://doi.org/10.1128/jvi.01248-09>

-
75. Miorin L, Kehrer T, Sanchez-Aparicio MT, Zhang K, Cohen P, Patel RS, Cupic A, Makio T, Mei M, Moreno E, Danziger O, White KM, Rathnasinghe R, Uccellini M, Gao S, Aydillo T, Mena I, Yin X, Martin-Sancho L, Krogan NJ, . . . García-Sastre A (2020) SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling. *Proc Nat Acad Sci U S A*, 202016650. Advance online publication. <https://doi.org/10.1073/pnas.2016650117>



miRNA Target Prediction: Overview and Applications

14

Fazlur Rahman, Sajjadul Kadir Akand, Muniba Faiza, Shams Tabrez, and Abdur Rub

Abstract

MicroRNAs (miRNAs) are short endogenous (~ 22 nucleotides long) noncoding RNAs synthesized by RNA polymerase class II enzyme in the nucleus. The exploration of miRNAs has become an emerging field of research due to their epigenetic regulation associated with a wide array of human diseases including cholera, hepatitis, malaria, and leishmaniasis. The miRNAs that are involved in a disease can be used as a biomarker due to their upregulated or downregulated expression level. miRNAs regulate the expression of mRNA through complementary base pairing with its 3-prime untranslated regions (3' UTRs). miRNAs are categorized according to its processes of precise formation demarcated as the canonical and noncanonical biogenic pathway. Since a huge amount of mRNA and miRNA data have been generated from the past researches, computational methods are needed to provide experimental validation with statistically significant outcomes. Computational approach is considered as one of the robust methods for miRNA target prediction.

In this chapter, we have focused our discussion on different miRNA prediction tools and their features in detail along with its role in infectious diseases.

Keywords

miRNA · RNA · SiRNA · Target prediction · Tools · 3'-UTR

F. Rahman · S. K. Akand · M. Faiza · S. Tabrez · A. Rub (✉)
Infection and Immunity Lab (414), Department of Biotechnology, Jamia Millia Islamia (A Central University), New Delhi, India
e-mail: arub@jmi.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

241

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*,
https://doi.org/10.1007/978-981-16-0691-5_14

14.1 Introduction

The world of RNA and its species has widened the scope of scientific research after the two decades of the discovery of various subtypes of nonprotein-coding RNAs (ncRNAs) [1, 2]; long nonprotein-coding RNAs (lncRNAs) [3, 4], small interfering RNAs (siRNAs) [5, 6], piwiRNA [7], circular RNAs (circRNAs) [8], or microRNAs (miRNAs) [9, 10]. miRNA was discovered in 1993 in a nematode worm, *Caenorhabditis elegans* [11]. Since then, it has been reported in viruses, single-celled eukaryotes, plants, and animals. miRNAs are 22–25-nucleotide-long endogenous noncoding RNAs. It regulates several biological processes through controlling the expression of target genes [12, 13]. Recently, the exploration of miRNAs has become pivotal for the detailed understanding of the alteration of genes in many human diseases [13]. miRNAs have been studied in context with several diseases including infectious as well as noninfectious [14–16].

A lot of research has been done in recent years to infer the miRNA targets and their functions which resulted in a huge amount of mRNA and miRNA data; therefore, computational methods provide experimental validation with statistically significant outcomes.

14.2 Formation of miRNA

miRNAs are synthesized by RNA polymerase class II enzyme in the cell nucleus. Initially, these are formed as a long primary transcript called pri-miRNA. These are formed either from the intronic regions of coding genes or from their noncoding regions of the genome. After its formation, pri-miRNA folds into a hairpin structure. miRNAs are categorized by precise formation processes demarcated as the canonical and noncanonical biogenic pathway (Fig. 14.1). Mostly pri-miRNA follows canonical biogenic pathway distinguished by double processing through the enzymes, namely, *Drosha* and *Dicer*, of the ribonuclease III (RNase III) family.

Drosha forms a microprocessor complex with DGCR8 in the nucleus [17–19]. This complex cuts the pri-miRNA and forms pre-miRNA. Exportin-5 helps in the transportation of pre-miRNA to the cytoplasm. Further, *Dicer* carries out the second cleavage of the pre-miRNA in the cytoplasm and forms ~21-nucleotide-long miRNA duplex with 2 nucleotide 3' overhang [19]. One of the mature strands from the duplex is laden onto an Argonaute (AGO) protein-associated RNA-induced silencing complex (RISC). This multiprotein complex leads to the formation of the effector complex in the cytoplasm. RISC uses miRNA as an antisense strand to recognize and regulate the specific targets. It is also reported that transposable elements are also involved in the miRNA formation [20]. Unlike animals, in plants, miRNA becomes fully mature in the nucleus [21, 22]. Recently, other processes of miRNA biogenesis have been categorized as noncanonical biogenic pathways consisting of *Dicer*-independent and *Drosha*-independent mechanism as explained in Fig. 14.1 [23–31].

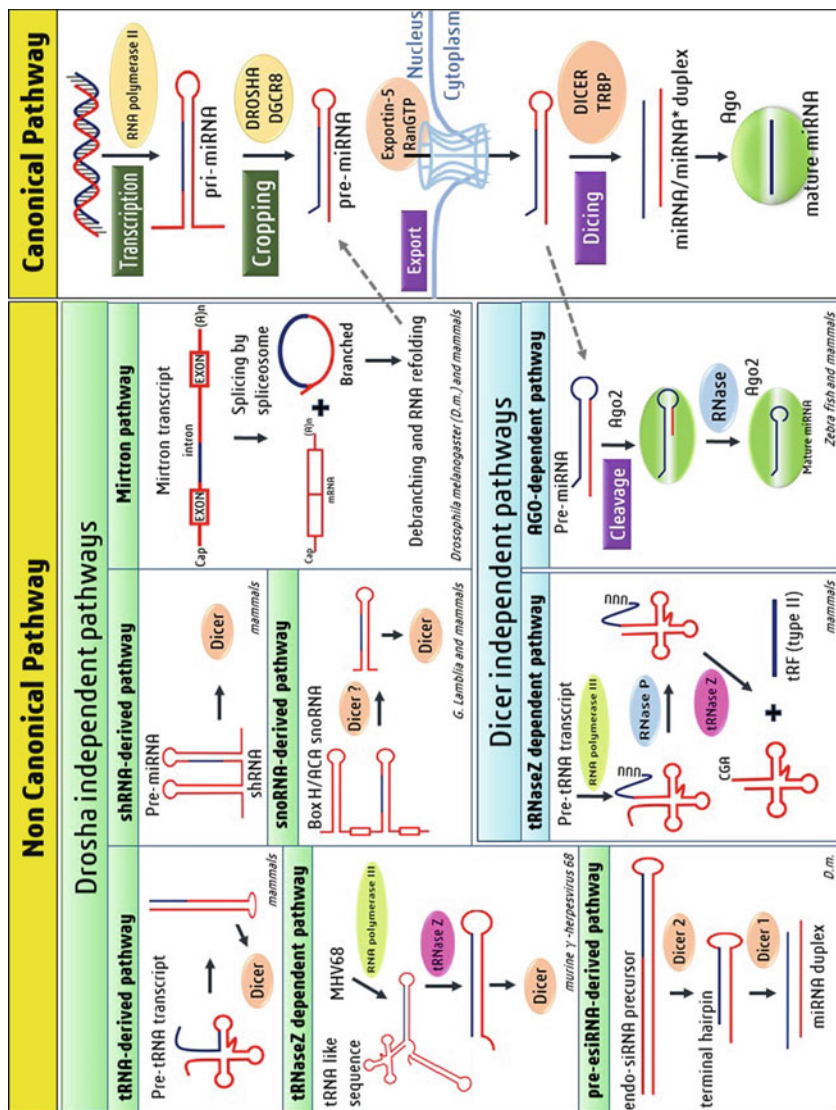


Fig. 14.1 Diagrammatic representation of noncanonical and canonical biogenic pathway of miRNA

14.3 Targets of miRNAs

miRNA regulates the expression of mRNA through binding with the 3' untranslated regions (UTRs) of mRNA through any of the two binding patterns as mentioned below [32]:

- (a) In first class, two to seven nucleotides from the 5' end of the miRNAs have complete Watson-Crick pairing to the 3' UTR of the mRNA. These nucleotide sequences are known as the seed region. Depending on the number of nucleotides involved in the base pairing, the seed regions are known as 6mer, 7mer, and 8mer accordingly.
- (b) While in another class of mRNA expression regulation by miRNA, there is partial base pairing formed in the seed region, but to overcome this insufficient binding, there are some base pairings found in the 3' side of the miRNA too.

A single mRNA may be regulated by multiple miRNAs and vice versa. Even one miRNA may influence the expression of other miRNAs, and due to this reason, regulatory mechanism becomes more complex.

14.4 Functions of miRNAs

miRNAs have a significant role in the modulation of gene expression in animals and in plants. In animals, miRNAs play important functions at various physiological and evolutionarily conserved developmental stages. miRNAs mostly show partial complementarity with the specific mRNAs in animals; however, it is adequate to regulate diverse physiological processes at the initiation step accompanied by the degeneration of mRNA [33]. In animals, miRNAs bind partially with the 3' UTR regulatory elements of the transcript and do not affect the functions of 5' UTR and target sites [34]. miRNAs affect the stability of the specific mRNAs at the transcriptional stage [35]. According to few studies, miRNAs can enhance translation in some specific cell types under certain conditions [36]. mRNAs seem to be inhibiting miRNAs from regulating the RNAs also. Like animals, plant miRNAs have shown key roles in the facilitation of organ maintenance at several developmental stages. miRNAs in plants have complementary regions to the specific mRNAs. In *Arabidopsis thaliana*, it has been shown that these regulates negatively also and may be due to their binding to the nonspecific targets in other noncoding RNAs [37].

14.5 miRNAs in Infectious Diseases

Different miRNAs are involved in a wide array of diseases, and the levels of miRNAs are either upregulated or downregulated based upon the disease. However, our focus centers around the miRNAs involved in the infectious diseases and their expression profiles. miR-93-5p belongs to the miR-106-25 cluster which plays a

crucial role in cancer development. HCV-1b core protein enhances the expression of miR-93-5p in Huh7 cells. IFNAR1 is a direct target of miR-93-5p, and upregulation of miR-93-5p inhibits IFN signaling pathway by reducing the phosphorylation status of STAT1, a transcription factor involved in IFN signaling [38]. During HBV, infection level of miR-29 is upregulated in HCC cell line HepG2.2.15. SMARCE1 is a target of miR-29a; an inhibition of SMARCE1 increased the HBV replication and expression [39]. The hepatic miR-122 level is reduced in HCV-infected cultured hepatocytes, and expression of NF- κ B-inducing kinase (NIK) is induced [40].

miR-146a plays an important role in immunomodulation of the host response during *Vibrio cholerae* infection. *V. cholerae* bacteria releases outer membrane vesicles which upregulate miR-146a. Upregulation of miR-146a allows the pathogen to colonize inside the host due to the reduced strength of an epithelial innate immune defense reaction [41]. miR-155 is another important miRNA that is induced by the *V. cholerae* to limit the host immune response, thereby reducing the probability of being eliminated. The miR-155 is upregulated in *V. cholerae* infection [42].

miR-155 acts as a negative regulator of endothelial and blood-brain barrier integrity during severe malaria. Upregulation of miR-155 is associated with increased endothelial activation and blood-brain barrier breakdown [43]. Severe malaria with multiorgan failure includes different miRNAs. Mouse with cerebral malaria shows upregulation of miR-27a, miR-150, and let7i levels in brain tissue compared to a mouse with no cerebral malaria [44]. miR-451 is a negative regulator of the host immune responses to *Plasmodium* infection. Downregulation of miR-451 induces pathogen clearance of *Plasmodium vivax* by CD4⁺ T cells [45]. Levels of plasma miR-451 and miR-16 were significantly downregulated in *Plasmodium vivax* infection, and these miRNAs can be exploited as biomarkers for malaria infection [46].

Autophagy happens to be one of the key events in the survival mechanism of *Leishmania* parasite under a stressed condition. The autophagy-related gene *BECN1* is a key autophagy-promoting gene. Upon *L. donovani* infection, the miR-30 family member, miR-30A-3p, modulates the autophagy by targeting *BECN1*. miR-30A-3p targets *BECN1* and decreases its level of expression resulting in decreased autophagic activity and finally parasite elimination [47]. miR-122 is expressed abundantly in the liver and modulates a wide range of liver functions. The *L. donovani* metalloprotease gp63 targets the DICER1 in human hepatocytes to reduce miR-122 expression [48]. Dendritic cells and macrophages infected with *L. donovani* show upregulation of let-7a and let-7b miRNAs, while these miRNAs are downregulated in *L. major*-infected cells. Suppressor of cytokine signaling 4 (SOCS4), a negative regulator of JAK-STAT signaling, plays the role in the regulation of miRNA expression [49]. *L. donovani* significantly downregulates the expression of miR-494 in THP-1-differentiated human macrophages. *L. donovani* metalloprotease gp63 degrades c-Jun and specifically upregulates the expression of Rab5a to downregulate miR-494 [50].

14.6 Fundamental Features Considered for miRNA Target Prediction

miRNAs target the complementary sequences generally present in the 3' UTR of mRNA. miRNA reduces the translation of mRNA to prevent protein synthesis. There are several computational methods which provide support to predict the possibilities and specificity of a miRNA binding to an mRNA. miRNA target prediction tools are based on a few common features [51], which are the basic features that are often implemented to develop algorithms for the miRNA tools. These features are described as follows.

14.6.1 Seedmatch

The initial two to eight nucleotides from 5' end to 3' end are called as seed sequence. Generally, miRNA target prediction tools are based on the Watson-Crick base pairing between the seed region of miRNA and mRNA. An ideal seed match has no gap in alignment of this region. There are different types of seed sequences that are used in the target prediction tool of miRNA [52]. Commonly used seed matching in miRNA tools are mentioned below:

1. *6-mer*: It includes the complementary pairing between the second nucleotide of the 5' side of the miRNA to the sixth nucleotide with the 3' UTR of the miRNA.
2. *7-mer-m8*: This is an ideal seed match between the second nucleotide from the 5' side of the miRNA to the eighth nucleotide with the 3' UTR of the miRNA
3. *7mer-A1*: This also is a perfect seed match between the second nucleotide from the 5' side of the miRNA to the seventh nucleotide with the 3' UTR of the miRNA in addition to an A across the miRNA first nucleotide.
4. *8-mer*: This type of seed match is also a perfect seed match between the second nucleotide from the 5' side of the miRNA to the eighth nucleotide with the 3' UTR of the miRNA in addition to an A across the miRNA first nucleotide.

14.6.2 Conservation

Conservation is defined as the measure of the sequence conserved throughout the different species. The functionality of an identified miRNA target is predicted by conservation analysis [51]. This feature is considered for the prediction of miRNA, 3' UTR, and 5' UTR and the combination of these three. The seed region in miRNA is considered to be more conserved than any other regions. The 3'-compensatory sites are a small portion of miRNA which has conserved pairing and networks with the target mRNA to compensate the mismatched seed [53].

14.6.3 Free Energy

The stability of a particular interaction-complex is defined by Gibb's free energy and is calculated as the variation in free energy change (ΔG). The value of ΔG suggests the stability of a reaction. A larger value of negative ΔG signifies greater stability of the reaction. miRNA binding with a target mRNA leads to the formation of either a stable structure or an unstable one. Formation of most stable interaction-complex is the root base of prediction of the likely target of miRNA.

14.6.4 Site Accessibility

Site accessibility is the degree of easiness through which a miRNA may work out its target mRNA and interacts with it. Interaction between miRNA and its target mRNA takes place in two steps. First, miRNA interacts with a short accessible part of mRNA, and then unfolding of mRNA takes place. After unfolding of mRNA, binding of miRNA occurs with it [54]. The amount of energy required for making a site in the mRNA accessible for binding of miRNA is required for the determination of site accessibility. Therefore, site accessibility is one of the most common features to find the appropriate target of a miRNA.

14.6.5 Other Features

There are some other features including target-site abundance, local AU content, GU wobble seed match, 3'-compensatory pairing, seed pairing stability, and position contribution that are often considered in miRNA target prediction algorithms.

14.7 Common Tools for miRNA Target Prediction

There are many prediction tools reported for the prediction of the miRNA targets. These all are based on the above-discussed features. Some of the tools are discussed in this section (Fig. 14.2) [55].

14.7.1 miRanda

It is a target prediction tool of miRNAs in a genome. The algorithm of this tool had been developed in C language. It was developed by the Computational Biology Center, at Memorial Sloan Kettering Cancer Center, USA, and is freely available online. Its algorithm works in the following three steps:

1. A miRNA sequence is entered as an input and is searched for the WC matches against the 3' UTR provided by the user.

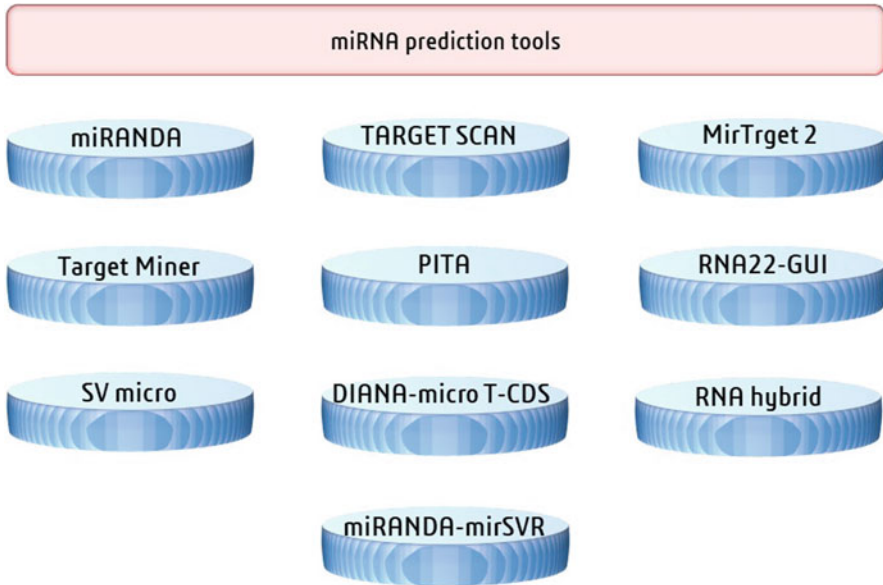


Fig. 14.2 An overview of the prediction tools

2. The free energy is determined for every miRNA-mRNA target pair, only when the pair surpasses a threshold that matches the score.
3. It applies the conservation analysis and is based on either the score or by having many sites.

14.7.2 miRanda-mirSVR

It is an online tool that recognizes candidate target sites and then scores them. SVR stands for super vector regression that is much similar to SVM (support vector machine). It comprises the study of AU flanking, site accessibility, and target-site position in the 3' UTR region. It also offers the miRNA expression analysis and links to the miRBase and miRO [56].

14.7.3 TargetScan

TargetScan prediction tool permits the operator to search the target by the name of the gene. The targets are predicted by calculating the predicted effectiveness of targeting, i.e., context + score, or by calculating the likelihood of conserved targeting, known as P_{CT} . P_{CT} is the probability of a target being targeted efficiently. No sequence is required to be input for the prediction. For conservation analysis, 3' UTR conservation is evaluated, and then specific k-mer is analyzed [57].

14.7.4 DIANA-microT-CDS

It is the latest version of DIANA-microT, first target prediction tool in human [58]. It incorporates data in a way to identify the most related features retrieved from photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) that allows this tool to get an idea about the position of miRNA interaction in the coding sequences and in 3' UTR region. There are numerous sites in a target that are recognized using microarray expression data. It gives the predicted target location, conservation, score, and external links to miRBase, Ensembl, and PubMed [58].

14.7.5 MirTarget2

MirTarget2 uses SVM and incorporates a large microarray training dataset in the target search. Its algorithm finds out the seed conservation, base content in the flanking regions, seed match in positions of 2–8 nucleotides, the region of target site in 3' UTR, and secondary structure [59, 60]. It has a disadvantage that it includes 3' UTR sequences as a training dataset with only one seed pairing site rather than many sites. It is comprehensible, and miRNA-mRNA interactions can be investigated by either mRNA or miRNA. Since there may be multiple target sites of miRNA in the target mRNA, therefore, there is a target running option that helps to select some specific target sites [59, 60].

14.7.6 RNA22-GUI

RNA22-GUI is the latest version of RNA22 [61, 62]. It identifies the target islands using pattern discovery and calculation of free energy of these islands and then finally predicts the candidate miRNA [63].

14.7.7 TargetMiner

TargetMiner identifies the seed sites between the user-provided miRNA and mRNA using an SVM-based classifier. This tool incorporates negative and positive training data to predict more precise seed matches in miRNA and the mRNA [64].

14.7.8 SVMicrO

SVMicrO is also a machine learning tool that includes a large positive dataset of different species [65]. It identifies the miRNA-miRNA interactions involving features of the miRNA target prediction including seed match, site accessibility, conservation, target-site abundance, and free energy [65].

14.7.9 Probability of Interaction by Target Accessibility (PITA)

PITA predicts the miRNA target using the target accessibility as a major feature [66]. It is dependent on the key observation that highly accessible regions of the 3' UTR of mRNAs have privileged and conserved positioning of target sites. This tool first recognizes a probable site through seed match measures. Then it considers site availability by calculating the free energy score depending on the difference between the gain of free energy linked with miRNA-mRNA hybrid target structure and the free energy requirement of dissociating the target to render it reachable. Further, to determine the interaction score for the microRNA and 3' UTR, the target-site abundance is considered which is computed through accessibility scores of the same miRNA [66].

14.7.10 RNAhybrid

RNAhybrid has a user-defined seed region and calculates the free energy between miRNA and mRNA. It predicts the target sites by assigning a p-value to the miRNA-mRNA interaction sites [67].

Among all the above-discussed tools miRanda is the most widely accepted tool for miRNA target prediction and also according to the ease of use [51, 68, 69]. But other tools such as RNA22 and GUI offer a graphical representation of the interactions between the miRNA and the mRNA.

References

1. Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. *Nat Rev Genet* 10:94–108
2. Rueda A, Barturen G, Lebrón R, Gómez-Martín C, Alganza Á, Oliver JL, Hackenberg M (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res* 43:W467–W473
3. Khvorovova A, Reynolds A, Jayasena SD (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell* 115:209–216
4. Ro S, Park C, Young D, Sanders KM, Yan W (2007) Tissue-dependent paired expression of miRNAs. *Nucleic Acids Res* 35:5944–5953
5. Ning S, Zhang J, Wang P, Zhi H, Wang J, Liu Y, Gao Y, Guo M, Yue M, Wang L, Li X (2016) Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res* 44:D980–D985
6. Spielmann N, Wong DT (2011) Saliva: diagnostics and therapeutic perspectives. *Oral Dis* 17:345–354
7. Flintoft L (2013) Non-coding RNA: structure and function for lncRNAs. *Nat Rev Genet* 14:598
8. Mittal V (2004) Improving the efficiency of RNA interference in mammals. *Nat Rev Genet* 5:355–365
9. Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, Kadener S (2014) circRNA biogenesis competes with pre-mRNA splicing. *Mol Cell* 56:55–66

10. Bahn JH, Zhang Q, Li F, Chan TM, Lin X, Kim Y, Wong DT, Xiao X (2015) The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva. *Clin Chem* 61:221–230
11. Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75:843–854
12. He L, Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* 5:522–531
13. Liu ZP, Wu H, Zhu J, Miao H (2014) Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza a virus infection. *BMC Bioinform* 15:336
14. Brase JC, Johannes M, Schlomm T, Fälth M, Haese A, Steuber T, Beissbarth T, Kuner R, Sültmann H (2011) Circulating miRNAs are correlated with tumor progression in prostate cancer. *Int J Cancer* 128:608–616
15. Srivastava K, Srivastava A (2012) Comprehensive review of genetic association studies and meta-analyses on miRNA polymorphisms and cancer risk. *PLoS One* 7:e50966
16. Yang H, Kong W, He L, Zhao JJ, O'Donnell JD, Wang J, Wenham RM, Coppola D, Kruk PA, Nicosia SV, Cheng JQ (2008) MicroRNA expression profiling in human ovarian cancer: miR-214 induces cell survival and cisplatin resistance by targeting PTEN. *Cancer Res* 68:425–433
17. Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the microprocessor complex. *Nature* 432:231–235
18. Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, Shiekhattar R (2004) The microprocessor complex mediates the genesis of microRNAs. *Nature* 432:235–240
19. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S, Kim VN (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425:415–419
20. Smalheiser NR, Torvik VI (2005) Mammalian microRNAs derived from genomic repeats. *Trends Genet* 21:322–326
21. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136:215–233
22. Park MY, Wu G, Gonzalez-Sulser A, Vaucheret H, Poethig RS (2005) Nuclear processing and export of microRNAs in Arabidopsis. *Proc Natl Acad Sci U S A* 102:3691–3696
23. Babiarz JE, Ruby JG, Wang Y, Bartel DP, Blelloch R (2008) Mouse ES cells express endogenous shRNAs, siRNAs, and other microprocessor-independent, dicer-dependent small RNAs. *Genes Dev* 22:2773–2785
24. Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC (2007) Mammalian mirtron genes. *Mol Cell* 28:328–336
25. Cheloufi S, Dos Santos CO, Chong MM, Hannon GJ (2010) A dicer-independent miRNA biogenesis pathway that requires ago catalysis. *Nature* 465:584–589
26. Cifuentes D, Xue H, Taylor DW, Patnode H, Mishima Y, Cheloufi S, Ma E, Mane S, Hannon GJ, Lawson ND, Wolfe SA, Giraldez AJ (2010) A novel miRNA processing pathway independent of dicer requires Argonaute2 catalytic activity. *Science* 328:1694–1698
27. Flynt AS, Greimann JC, Chung WJ, Lima CD, Lai EC (2010) MicroRNA biogenesis via splicing and exosome-mediated trimming in Drosophila. *Mol Cell* 38:900–907
28. Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC (2007) The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* 130:89–100
29. Okamura K, Lai EC (2008) Endogenous small interfering RNAs in animals. *Nat Rev Mol Cell Biol* 9:673–678
30. Ruby JG, Jan CH, Bartel DP (2007) Intronic microRNA precursors that bypass Drosha processing. *Nature* 448:83–86
31. Yang JS, Lai EC (2011) Alternative miRNA biogenesis pathways and the interpretation of core miRNA pathway mutants. *Mol Cell* 43:892–903
32. Saito T, Saetrom P (2010) MicroRNAs--targeting and target prediction. *New Biotechnol* 27:243–249
33. Kato M, Slack FJ (2008) microRNAs: small molecules with big roles – *C. elegans* to human cancer. *Biol Cell* 100:71–81

34. Guo H, Ingolia NT, Weissman JS, Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466:835–840
35. Fabian MR, Sonenberg N, Filipowicz W (2010) Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem* 79:351–379
36. Vasudevan S, Tong Y, Steitz JA (2007) Switching from repression to activation: microRNAs can up-regulate translation. *Science* 318:1931–1934
37. Chitwood DH, Timmermans MC (2007) Target mimics modulate miRNAs. *Nat Genet* 39:935–936
38. He CL, Liu M, Tan ZX, Hu YJ, Zhang QY, Kuang XM, Kong WL, Mao Q (2018) Hepatitis C virus core protein-induced miR-93-5p up-regulation inhibits interferon signaling pathway by targeting IFNAR1. *World J Gastroenterol* 24:226–236
39. Wu HJ, Zhuo Y, Zhou YC, Wang XW, Wang YP, Si CY, Wang XH (2017) miR-29a promotes hepatitis B virus replication and expression by targeting SMARCE1 in hepatoma carcinoma. *World J Gastroenterol* 23:4569–4578
40. Lowey B, Hertz L, Chiu S, Valdez K, Li Q, Liang TJ (2019) Hepatitis C virus infection induces hepatic expression of NF- κ B-inducing kinase and lipogenesis by downregulating miR-122. *mBio* 10:e01617-19
41. Bitar A, Aung KM, Wai SN, Hammarström ML (2019) *Vibrio cholerae* derived outer membrane vesicles modulate the inflammatory response of human intestinal epithelial cells by inducing microRNA-146a. *Sci Rep* 9:7212
42. Bitar A, De R, Melgar S, Aung KM, Rahman A, Qadri F, Wai SN, Shirin T, Hammarström ML (2017) Induction of immunomodulatory miR-146a and miR-155 in small intestinal epithelium of *Vibrio cholerae* infected patients at acute stage of cholera. *PLoS One* 12:e0173817
43. Barker KR, Lu Z, Kim H, Zheng Y, Chen J, Conroy AL, Hawkes M, Cheng HS, Njock MS, Fish JE, Harlan JM, López JA, Liles WC, Kain KC (2017) miR-155 modifies inflammation, endothelial activation and blood-brain barrier dysfunction in cerebral malaria. *Mol Med* 23:24–33
44. Chamnanchanunt S, Fucharoen S, Umemura T (2017) Circulating microRNAs in malaria infection: bench to bedside. *Malar J* 16:334
45. Chapman LM, Ture SK, Field DJ, Morrell CN (2017) miR-451 limits CD4(+) T cell proliferative responses to infection in mice. *Immunol Res* 65:828–840
46. Chamnanchanunt S, Kuroki C, Desakorn V, Enomoto M, Thanachartwet V, Sahassananda D, Sattabongkot J, Jenwithisuk R, Fucharoen S, Svasti S, Umemura T (2015) Downregulation of plasma miR-451 and miR-16 in *Plasmodium vivax* infection. *Exp Parasitol* 155:19–25
47. Singh AK, Pandey RK, Shaha C, Madhubala R (2016) MicroRNA expression profiling of *Leishmania donovani*-infected host cells uncovers the regulatory role of MIR30A-3p in host autophagy. *Autophagy* 12:1817–1831
48. Ghosh J, Bose M, Roy S, Bhattacharyya SN (2013) *Leishmania donovani* targets Dicer1 to downregulate miR-122, lower serum cholesterol, and facilitate murine liver infection. *Cell Host Microbe* 13:277–288
49. Geraci NS, Tan JC, McDowell MA (2015) Characterization of microRNA expression profiles in *Leishmania*-infected human phagocytes. *Parasite Immunol* 37:43–51
50. Verma JK, Rastogi R, Mukhopadhyay A (2017) *Leishmania donovani* resides in modified early endosomes by upregulating Rab5a expression via the downregulation of miR-494. *PLoS Pathog* 13:e1006459
51. Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L, Congdon CB (2014) Common features of microRNA target prediction tools. *Front Genet* 5:23
52. Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD (2019) A practical guide to miRNA target prediction. *Methods Mol Biol* 1970:1–13
53. Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 19:92–105
54. Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y (2007) Potent effect of target structure on microRNA function. *Nat Struct Mol Biol* 14:287–294

55. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human MicroRNA targets. *PLoS Biol* 2:e363
56. Betel D, Koppal A, Agius P, Sander C, Leslie C (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* 11:R90
57. Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *elife* 4:e05005
58. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res* 41:W169–W173
59. Wang X (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* 14:1012–1017
60. Wang X, El Naqa IM (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics* 24:325–332
61. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. *Monatshfte für Chemie/Chemical Monthly* 125:167–188
62. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 126:1203–1217
63. Loher P, Rigoutsos I (2012) Interactive exploration of RNA22 microRNA target predictions. *Bioinformatics* 28:3322–3323
64. Bandyopadhyay S, Mitra R (2009) TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics* 25:2625–2631
65. Liu H, Yue D, Chen Y, Gao SJ, Huang Y (2010) Improving performance of mammalian microRNA target prediction. *BMC Bioinform* 11:476
66. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. *Nat Genet* 39:1278–1284
67. Krüger J, Rehmsmeier M (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res* 34:W451–W454
68. Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD (2016) Bioinformatic tools for microRNA dissection. *Nucleic Acids Res* 44:24–44
69. Faiza M, Tanveer K, Fatihi S, Wang Y, Raza K (2019) Comprehensive overview and assessment of microRNA target prediction tools in homo sapiens and *drosophila melanogaster*. *Curr Bioinforma* 14:432–445

Part III

**Infection Proteomics: Proteomics Applications
for Human Pathogens**



V. S. Gowri and V. Sabareesh

Abstract

Rapid growth in the emergence of several microbial strains that show resistance towards a variety of antibiotics and other drugs is posing a serious issue, all over the world. Peptide-based compounds have been sought as alternative agents to resolve this global crisis. The key reason being several peptides, both natural and engineered analogues (motivated from natural ones), were found to elicit good potencies even against the drug-resistant strains. Peptidic molecules were found to have therapeutic potential toward other human diseases as well, such as cancer and diabetes. Consequently, the need for high-throughput techniques that can facilitate swift identification and characterization of peptides from diverse natural sources was imperative. Chromatography, in particular high-performance liquid chromatography (HPLC) and mass spectrometry-based methods, could be adapted for high-throughput characterization of natural peptides. In order to process and analyze huge data sets that result from such high-throughput experiments, development of computational tools, particularly databases, was also necessary. Eventually a new branch of study called “peptidomics” emerged, which encompasses various experimental and computational strategies. A few aspects of this “omics” approach, mainly focusing on sequencing of peptides from diverse biological sources, are described in this chapter.

V. S. Gowri

PG & Research Department of Chemistry, Auxilium College (Autonomous), Vellore, Tamil Nadu, India

V. Sabareesh (✉)

Centre for Bio-Separation Technology (CBST), Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India

e-mail: v.sabareesh@vit.ac.in

Keywords

Natural products · Therapeutic peptides · Peptidomics · High-throughput · Databases

15.1 Introduction

Peptide-based natural products constitute an important class of drugs, which have been discovered and developed from a variety of natural or biological sources such as bacteria, fungi, plants, and animals [1–3]. Depending on the target and the level of potency, many a times, the natural peptides are suitably engineered and transformed to synthetic or semi-synthetic analogues, so as to enhance their function and render them more potent than the naturally observed level of activity [4, 5]. Further, attempts to modify the natural peptides are also known, in order to decrease and alter their natural toxicity, which can improve their therapeutic efficacy. In other words, design of many of the synthetic peptide-based drug molecules is largely motivated from the functions observed with the natural peptides. And, the enhancement of the function can be achieved by suitably altering the molecular structure of the peptide (ligand), since it is well-known that structure dictates the activity on the target, i.e., structure-activity relationship (SAR). Thus, the process of drug discovery would involve two major steps: (1) to conduct different types of bioactivity assays for identifying target(s) and (2) to elucidate the sequence and structure of “bioactive” peptides.

15.1.1 Biosynthesis and Molecular Structural Properties of Peptides

Peptides are biosynthesized not only by conventional ribosomal routes but also by non-ribosomal synthesis, wherein the products resulting from the latter route are referred as secondary metabolites. In the case of conventional ribosomal synthesis, the peptides mature from a precursor protein (called as “prepropeptide precursor”), upon action of different proteases, in which the prepropeptide precursor is coded by a particular gene [6, 7]. Subsequently, posttranslational modifications (PTMs) might take place on such matured peptides, and such peptides are called RiPPs, meaning ribosomally synthesized and posttranslationally modified peptides [8]. Hydroxyproline (Hyp), disulfide bonds, and pyroglutamic acid are some examples of PTMs that are found in several peptides [9–11]. However, there are no genes that directly code for non-ribosomal peptides. The synthesis of these peptides involves non-ribosomal peptide synthetases (NRPSs). NRPSs are multienzyme complexes having modular architecture, wherein each module catalyzes a particular function [12, 13]. Therefore, non-ribosomal peptides contain non-proteinogenic amino acids, which are not usually found in proteins. α -aminoisobutyric acid (Aib), isovaleric acid (isovaline, Iva), α -aminobutyric acid (Abu), and ornithine (Orn) are a few examples of non-proteinogenic amino acids [13–15]. Table 15.1 summarizes

Table 15.1 A few examples of RiPPs and NRPs classified based on their architecture

S. no.	Peptide architecture	Source	References
RiPPs			
I	Linear peptides		
	Glucagon	Human	[6]
	Microcin B17	<i>Escherichia coli</i> (<i>E. coli</i>)	[16, 17]
II	Cyclic peptides		
a.	Head-to-tail cyclized only Cyclolinopeptide F—orbitide	Plant (linseed)	[18]
b.	Sidechain cyclized only Ziconotide (Prialt)— <i>Conus</i> peptide	<i>Conus magus</i> Marine cone snail	[19]
c.	Both head-to-tail and sidechain cyclized Kalata B1—cyclotide	Plant <i>Oldenlandia affinis</i>	[20, 21]
III	Peptides having ring/knotted structures with isopeptide bond		
	Microcin J25—lasso peptide (class II)	Bacteria (<i>E. coli</i> AY25)	[22, 23]
Non-ribosomal peptides (NRPs)			
I	Linear peptides		
a.	Alamethicin—peptaibol	Fungi (<i>Trichoderma viride</i>)	[24, 25]
b.	Gramicidin A	Bacteria	[26, 27]
II	Cyclic peptides		
a.	Gramicidin S	Bacteria	[28]
b.	Hymenamamide F	Marine sponge (<i>Hymeniacidon</i> sp.)	[29]
III	Cyclic depsipeptides		
a.	Surfactins	Bacteria	[30, 31]
b.	Enniatin, beauvericin	Fungi	[8, 32]
c.	Fengycins	Bacterial species including endophyte	[31, 33]
d.	Kahalalide F	Marine mollusc (<i>Elysia rufescens</i>)	[34]

some popularly known RiPPs and non-ribosomal peptides along with their respective biological sources of origin.

Another interesting feature is that ring-like structures are found in both these classes of peptides, due to cyclization involving backbone or side chain atoms (Fig. 15.1). But, the mechanism of cyclization through ribosomal biosynthetic pathway is distinctly different from that of non-ribosomal route [8, 13, 35]. An example of “backbone-cyclized” RiPPs is cyclolinopeptides belonging to “orbitides” class, which lacks disulfide bonds [18, 36] (Fig. 15.1). The “backbone-cyclized” peptides are also known as head-to-tail-cyclized peptides [8]. Those peptides that are not backbone-cyclized, but possess disulfide bonds only, can be called as “sidechain-cyclized” peptides, wherein a cyclic structure is formed, when the “sidechain” sulfur atoms of two cysteine amino acids in a peptide are connected giving rise to a “disulfide bond (intramolecular)”.

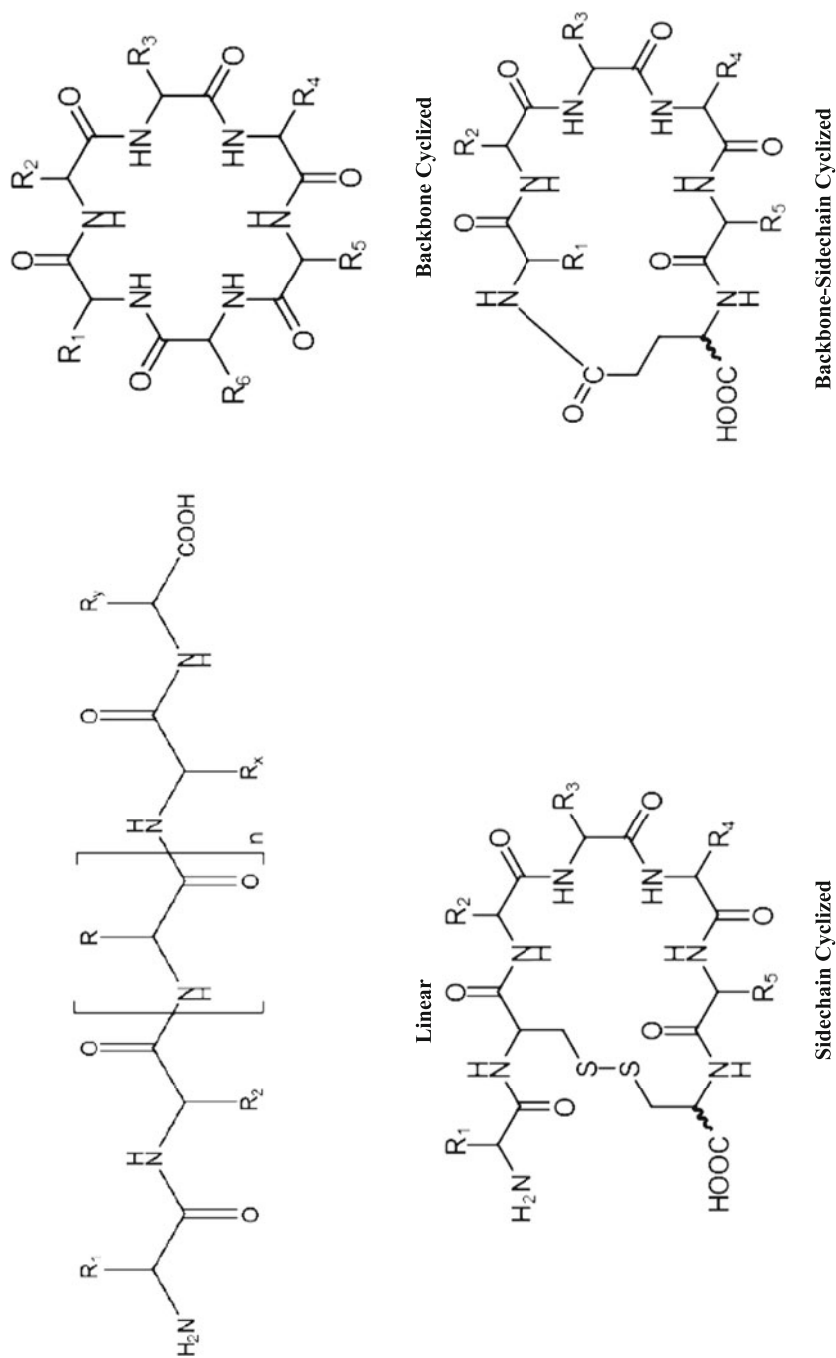


Fig. 15.1 Classification of peptides based on their architecture due to different connectivity between backbone and sidechain atoms. Note: Stereochemistry is not shown

Conotoxins and defensins are examples of sidechain-cyclized peptides, as they possess only disulfide bonds [11, 37]. Interestingly, “cyclotides” from plants are backbone and sidechain cyclized, and therefore, they form knot-like structures [35, 36]. Kalata B1 and cycloviolacins are some well-studied examples of cyclotides, which are in fact RiPPs [20, 21]. Cyclodepsipeptides are an interesting class of NRPs in that they have at least one ester bond amid other amide bonds. These cyclic peptides have been discovered in both fungi and bacteria. While enniatins and beauvericins are of fungal origin, surfactins and fengycins have been identified in bacterial species. Such peptides are known from marine sources as well. With respect to stereochemistry, D-amino acids containing natural peptides are well-known [38], though the preponderance of L-amino acids is higher than the D-amino acids in many natural peptides.

15.1.2 Omics Approach to Investigate Natural Peptides

In the recent past, particularly during the last decade, *omics* approach has become very successful not only to investigate genes and proteins [39] but also for studying peptides, lipids, carbohydrates, and other metabolites. An omics study simply means large-scale investigation of numerous samples. Therefore, the main criterion of an omics approach is that it should involve “high-throughput” analysis of the samples as well as high-throughput characterization of the molecules in those samples. Quite similar to the growth trend observed in the area of proteomics, the omics approach could be applied successfully for peptides also, primarily due to the advancements in the field of “biomolecular mass spectrometry,” not to forget the developments in the field of chromatography as well. In other words, development of different types of mass spectrometric methods, in conjunction with suitable chromatographic techniques, led to the emergence of the “peptidomics” field, which concerns with the high-throughput analysis of natural peptides from diverse biological sources [10, 40].

In a typical omics approach-based study, vast sets of experimental data would be generated, due to high-throughput analysis. Manual analysis of such huge data would be tedious. Consequently, computational methods are applied with the objective to simplify data processing and facilitate rapid analysis. One of the computational strategies is to construct “database”, which can be allowed to interrogate with the experimental data. The output resulting from such interrogative exercises can then be taken up for further interpretations. The database is designed according to the nature of experimental data and based on the requirements of a particular project. Thus, for peptidomic investigations, mass spectrometry, chromatography, and computationally built databases are the some essential components, among others. Various aspects related to omics approach, mass spectrometry, and databases are described in the following sections.

15.2 Mass Spectrometry and Omics Approach

15.2.1 Peptide Sequencing and Characterization: Traditional Methods

Sequencing of peptides and proteins, viz., elucidating the “primary structure,” was classically done by 2,4-dinitrofluoro benzene (FDNB). FDNB was used to deduce the primary structure of insulin, which is actually the first polypeptide to be fully sequenced, for which Frederick Sanger was awarded the 1958 Nobel Prize in Chemistry. FDNB is also called as Sanger’s reagent. However, later, Pehr Edman introduced another reagent, phenyl isothiocyanate (PITC), which was possible to be utilized for “automated” sequencing purposes [41]. Therefore, PITC eventually became a more preferable choice of reagent. Sequencing using PITC is also popularly known as N-terminal sequencing or Edman’s degradation method. The foremost criterion for N-terminal sequencing using PITC (and also in the case of FDNB) is that the peptide or protein must not possess blocked or modified N-terminus, such as acetyl or any other acyl modification (e.g., palmitoyl, myristoyl, etc.) or pyroglutamyl residue [9]. In other words, only those peptides and proteins that have “free” amino terminus can be sequenced by Edman’s method. This means that the cyclic peptides that are head-to-tail cyclized too cannot be sequenced by Edman’s method. Thus, in these cases, spectroscopic methods, e.g., nuclear magnetic resonance (NMR) spectroscopy, can be applied to elucidate the primary structure. X-ray diffraction or X-ray crystallography also can be useful to find the sequences of such N-terminus modified or head-to-tail cyclized peptides. However, prior to analysis by NMR spectroscopy and/or X-ray crystallography, the peptide must be isolated and purified from the crude biological extract. A biological extract would be concoction of numerous molecules including peptides, and hence, it must be subjected to multiple steps of sample processing until a purified peptide is isolated. In other words, spectroscopic techniques and X-ray crystallography are applicable on a pure sample that contains only the analyte molecule (peptide) of interest. Further, in the case of X-ray crystallography, a single crystal, which can give good quality diffraction data, must be obtained. Only then, it is possible to determine the three-dimensional molecular structure with fewer ambiguities. Therefore, the crystallization conditions need to be very carefully engineered.

15.2.1.1 Typical Protocol to Isolate and Purify Natural Peptides: A Brief Summary

The first and foremost step is to choose an appropriate biological species/source, which is selected based on the aim and objective of a particular project. According to the selected organism, one or a few particular parts (i.e., type of organ or tissue or cell) is/are chosen, e.g., root/stem/leaf of a plant or liver/pancreas/brain/skin tissue of an animal, etc. [10, 42–44]. In the case of microbial source, either the membrane or the intracellular compartment or even the whole cell is chosen, suiting to the objectives of a particular project. Depending on the nature of the selected specimen, different tools are employed for mechanical/physical processing and for

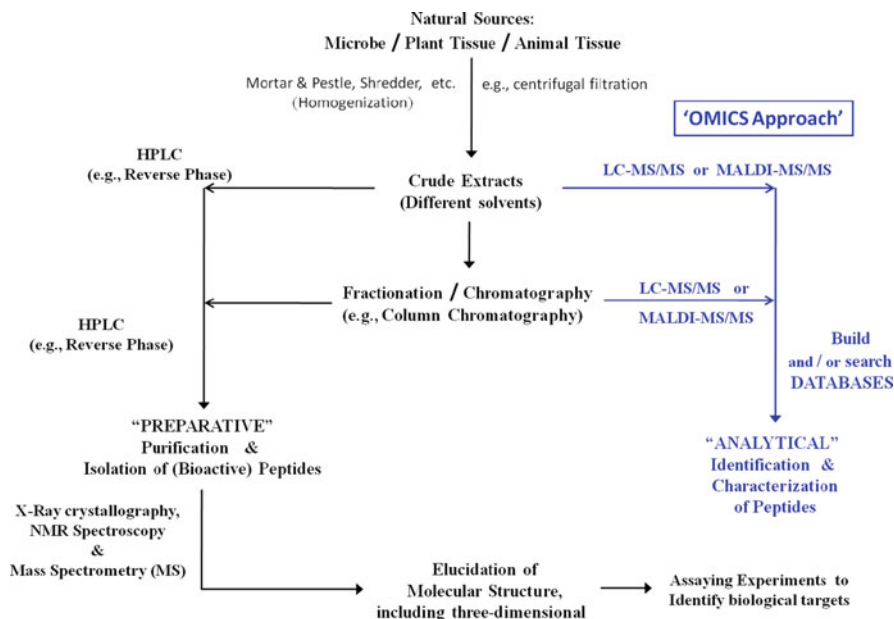


Fig. 15.2 A typical strategy for identification and characterization of peptides. The “omics approach” pertains to analytical detection of peptides directly from crude extracts or from semi/partially purified fractions. Here, LC mostly refers to reverse phase chromatography

homogenization, such as mincer, shredder, mortar & pestle, vortex machine, centrifugation, sonication, etc. [44]. Choice of solvents/buffers is also extremely vital, which can significantly influence the yield of the compounds in the extract [45, 46]. A variety of solvents or buffers (aqueous) or different combinations of solvents/buffers are utilized for optimizing the process of extraction. Based on the polar/non-polar nature, different solvent extracts contain different types of peptide molecules. Subsequently, the extract is subjected to pre-fractionation methods such as filtration, ultrafiltration, solid-phase extraction, etc. [14, 47]. Thereafter, several rounds of chromatography would be essential, until a purified peptide is isolated in good quantities or yield (see Fig. 15.2). Hence, selecting an appropriate chromatographic technique is another crucial step. Often more than one chromatographic method may have to be utilized not only to achieve sufficient yield but also to get the peptide with appreciable or high level of purity and eliminate as many contaminants as possible [9, 48]. With the advent of high-performance liquid chromatography (HPLC), the applications of different types of chromatography, in particular reverse phase (RP)-HPLC, soared up significantly [49, 50]. HPLC not only enabled fast rates of fractionation but also offered far better resolution and sensitivity than the traditional column chromatography [49, 50]. Fractionation by HPLC can be done in three different modes: (1) preparative, (2) semi-preparative, and (3) analytical. Depending on the quantity of the sample to be analyzed, one of the three modes needs to be chosen. Preparative and semi-preparative HPLC are suitable to isolate

purified product in larger quantities, approximately in the range of a few 100 milligrams to a few grams. Analytical HPLC is generally meant for getting a quick look at the total number of components in a particular extract, with lower sample consumption. Even then, analytical HPLC can be used for isolating compounds as well [46], but the yield of purified compound might be less than or equivalent to the quantities achieved with semi-preparative HPLC.

Further, it is also important to realize that the yield of the purified peptide is also dependent on the natural abundance of the peptide of interest existing in a particular biological extract, irrespective of the mode of HPLC. Moreover, the abundance of a particular peptide in an extract would also vary according to the nature of solvents used during the extraction procedure (*vide supra*). Indeed, various instrumental parameters of the HPLC equipment too can significantly influence the extent of purity and yield of the purified compound [50].

Several investigations have demonstrated the successful applications of RP-HPLC for isolation and purification of peptides, particularly peptide hormones and neuropeptides/neurotoxins, from diverse natural sources, such as brain tissue, scorpion venom, plants, marine organisms, soil microbes, etc. [42, 44–49, 51–53]. The separation of peptides by RP-HPLC is primarily based on the relative variations in the polar/apolar character of the constituent peptides in the biological extract, but it is ideally suited for hydrophobic (non-polar) peptides rather than hydrophilic. RP-HPLC involves hydrophobic or non-polar stationary phase (column) and polar mobile phase (solvents). Octadecyl (C18) silane-based material is often used in the stationary phase for RP-HPLC of peptides. C8 and C4 column chemistry also has been found to be fruitful in certain studies [9, 48, 54]. Polar solvents such as methanol, acetonitrile, isopropanol, and water are usually employed as mobile phase in the RP-HPLC. Ion-pairing agents, for instance, formic acid, acetic acid, trifluoroacetic acid, triethylamine, etc., are also necessary in several cases for enhancing the resolution of separation, since these reagents help in neutralizing the charges on the peptides, ensuring that separation on the RP column is predominantly due to polar/apolar nature only [39, 42, 45, 55]. The detection of eluted peptides in HPLC is usually accomplished by measuring absorbance in the ultraviolet (UV) region of electromagnetic radiation, typically 214 nanometers (nm) or 220 nm or 226 nm [46, 47, 52–54].

15.2.2 Why Mass Spectrometry?

Though direct analysis of crude extracts by NMR spectroscopic methods has been useful to a certain extent, particularly for metabolomic studies in recent times [56–58], its utility is limited for analyzing peptides within crude extracts. For X-ray crystallography, purity of the analyte molecule is a necessity to get single crystal. In certain (rare) cases, good diffraction quality crystals might yield from semi- or partially purified samples; however those crystals may not contain the molecule (peptide) of interest that is targeted/aimed for. High-throughput crystallization equipments are readily available to grow good-quality crystals quite rapidly

[59]. But, the processes involved in solving and determining the molecular structure after data collection on single crystals, not only is very time-intensive, but also requires appropriate knowledge and expertise. Thus, perceiving the number of steps implicated in isolating a purified peptide, it is apparent that NMR spectroscopy and X-ray crystallography are not suitable for high-throughput analysis or for high-throughput “screening” purposes of crude biological extracts, at least in the current era. However, mass spectrometry (MS) can be applied directly on crude biological concoctions, and it also can be adapted for high-throughput applications. Additionally, MS can be applied for molecular structural characterization of purified peptide samples as well, thereby complementing the NMR spectroscopic and X-ray crystallographic data. Nevertheless, NMR spectroscopy- and X-ray diffraction-based methods are extremely important and inevitable, for the determination of molecular structure and three-dimensional molecular conformation of the “purified” molecule (Fig. 15.2).

With regard to Edman’s degradation method, although the protocol could be automated for rapid sequencing of peptides and proteins possessing free N-terminus, it could be applied only on the purified peptide or protein only. So, the crude extracts cannot be directly characterized by Edman’s analysis. Thus, Edman’s method is not suited for high-throughput studies [41, 60]. However, with MS, it was possible to design and develop different methods and protocols that could deduce sequences of peptides in a high-throughput fashion [41, 61]. Therefore, it is essential to understand a few key fundamental principles of MS. The following section describes the basic principles of “biomolecular” MS and delineates some aspects that enable MS for high-throughput investigations.

15.2.2.1 Brief Introduction to Mass Spectrometry

MS is used to determine molecular mass and also to elucidate the molecular structure. “Ionization source” and “mass analyzer(s)” are two important parts of mass spectrometer. The purpose of ionization source is to generate “ions”, viz., charged molecular species or molecular ions in gas phase or vacuum. And the mass analyzer measures the mass and charge of the ions. The measurements made by the mass analyzer are plotted in the form of a graph showing the ratio of mass-to-charge (m/z) values versus intensity. Intensity refers to the number or population of the molecular ions, and such a graph is called “mass spectrum”. So, from the mass spectrum, it is possible to determine the population or abundance of ions having different m/z values.

Ionization Sources

Different mass spectrometric techniques are used for the analysis of different types of molecules. According to the relative polar or apolar character of the molecule (s) under investigation, a particular mass spectrometric technique/method is chosen. Especially, suitable ionization source needs to be employed, depending on the polar or non-polar property of the molecule. Electron ionization (EI) or electron impact ionization is particularly apt to investigate non-polar, volatile compounds. Hence, EI-MS is usually coupled with gas chromatography (GC), and it is referred as

GC-MS or specifically GC-EI-MS. EI produces radical ions, which have unpaired electron or which contain odd number of electrons. Radical ions are usually unstable, and depending on the site of formation of the radical (unpaired electron), the ion can undergo fragmentation in different ways. Because of the molecular fragmentation events upon electron impact in the source, the signal corresponding to the intact molecular ion (viz., intact molecular mass) is often detected with lower intensities, which is a characteristic feature of EI-MS. However, the fragmentation patterns have proven to be useful to deduce molecular structure and some important functional groups. Further, it is useful for analysis of small molecules, whose masses $< \sim 500$ Daltons (Da) only. Many peptides and proteins are, in fact, large-sized, non-volatile, and predominantly polar molecules. Hence, EI-MS or GC-MS is rather not ideal for characterizing peptides and proteins. Nevertheless, GC-EI-MS was successfully applied to elucidate the primary structure (i.e., sequence) of certain peptides and a few proteins, whereby different chemical derivatization methods were developed that could convert the polar, volatile nature of the amino acids and shorter peptides into apolar and non-volatile [41]. But, the strategy of combining chemical derivatization procedures along with GC-EI-MS to sequence peptides and proteins could not be extended for large-scale and routine analyses.

Since EI-MS cannot properly generate intact molecular ions due to significant molecular fragmentations, there were attempts to develop a new source with the aim to measure intact molecular mass by decreasing the extent of molecular fragmentation during the ionization process. Chemical ionization (CI) was another source, through which it was possible to generate intact molecular ions with little fragmentation. CI was built based on the principle of EI, whereby the electrons are allowed to interact with a “reagent gas”, e.g., ammonia or isobutane. The ions (including radical ions) generated from the reagent gas react with the analyte molecules, and results in the formation of intact molecular ions of the analyte. Hence, it was possible to obtain good signal intensity for intact molecular ion by CI. Thus, CI is relatively softer than EI. Observation of adduct ions, e.g., $[M + H]^+$ (where M denotes molecule and H denotes hydrogen or proton), is a typical attribute of CI. Further, it was possible to couple CI also with GC. Though CI proved to be useful to determine intact molecular masses of certain polar, non-volatile compounds, it was not possible to extend its application for very large-sized molecules. And thus, efforts to find newer methods of ionization continued, such as field ionization (FI), field desorption (FD), and fast atom bombardment (FAB). FI and FD ionization sources have been applied successfully to deduce the sequences of some peptaibols from fungal sources [62–64], albeit FI and FD were also not capable of analyzing large-sized molecules. However, development of FAB-MS enabled ionization of larger-sized polar, non-volatile molecules, thereby aiding in determining their intact molecular masses. In comparison to the earlier sources, the ionization process occurring during FAB was even softer, which helped in obtaining molecular ions in their whole/intact form, with very little fragmentation [65, 66]. So, FAB-MS found wider applications, particularly to study larger-sized molecules of masses up to ~ 6 kilo Daltons (kDa) or in some cases even ~ 10 kDa [67]. Another technique called “liquid secondary ion mass spectrometry (LSIMS)” was developed based on the principle of FAB, whereby

LSIMS could be used along with liquid chromatography [68]. For nearly a decade or so, FAB-MS was fruitful for analysis of several polar molecules, including peptides.

However, with the arrival of two new ionization methods: (1) electrospray ionization (ESI) and (2) matrix-assisted laser desorption and ionization (MALDI), the utility of FAB-MS plummeted, mainly because both MALDI and ESI were found to be more soft processes than FAB. MALDI and ESI could generate ions of “highly polar macromolecules”, viz., large-sized proteins of masses up to or greater than even ~100 kDa, in gas phase/vacuum, with very little dissociation [69–71]. Consequently, ESI- and MALDI-based MS proved to be of immense utility to investigate different types of peptides also, of diverse polarities. Another highlighting feature is that several water-soluble compounds, including glycans and polar lipids too, could be successfully characterized by these two ionization modes. In fact, the advent of ESI and MALDI facilitated “high-throughput” analyses of peptides, lipids, and glycans as well. Thus, these two ionization methods were largely responsible for the genesis of the **omics** approach not only for investigating proteins but also for other biomolecules aforesaid. Thus, new fields of study emerged, namely, proteomics, lipidomics, and glycomics. In recognition of their efforts for the development of ESI and MALDI, respectively, John Fenn and Koichi Tanaka shared one half of the Nobel Prize in Chemistry 2002 [72, 73]. Table 15.2 gives a brief summary of certain important aspects of different ionization sources.

ESI- and MALDI-MS can detect analytes of low (trace) concentrations in the picomole–nanomole range (viz., ~ nanograms) from low sample volumes, of the order of a few microliters. Moreover, these two ionization modes could be linked with liquid chromatography (LC) as well. In particular, reverse-phase LC could be successfully interfaced with ESI [75, 76], thereby giving rise to a new hyphenated technique, LC-MS or more specifically LC-ESI-MS. In fact, the development of such a hyphenated technique, which is also called as online LC-MS (analogous to GC-MS), enabled direct characterization of crude extracts, whereby the crude sample would be first introduced into the chromatographic column (typically reverse phase) and the eluents would be directed to the ionization source of the mass spectrometer [14, 75]. In other words, mass spectrometer became another detector for HPLC, and hence it was possible to acquire “mass-detected chromatogram” besides the typical UV chromatogram [75, 77–79]. And in recent times, hydrophilic interaction chromatography (HILIC) too could be coupled with ESI, viz., through LC-MS mode of operation, but this has been mostly for proteomic studies [80, 81]. In the case of MALDI, more samples can be introduced into the spectrometer using a sample plate consisting of 96 or 100 or 384 wells, where each well can accommodate 1 μ l or 2 μ l of sample.

Applications of LC-MALDI-MS are relatively less than LC-ESI-MS, though a few strategies have been devised to link LC with MALDI mass spectrometers [82, 83]. Further, the utility of LC-MALDI-MS is limited for peptidomic studies, when compared to proteomics [84, 85]. In fact, the rise of LC-MS-based methods involving ESI and MALDI may be regarded as the beginning of the “omics”

Table 15.2 Different ionization sources in mass spectrometry used for biomolecular analysis and some of their characteristics: a summary^a

S. no.	Ionization source	Nature of ions formed	Molecules that can be ionized	Approx. maximum limit of intact molecular mass detection
1.	Electron ionization (EI)	Radical ions: $M^{+\bullet}$ and $M^{-\bullet}$	Non-polar, volatile molecules	~ 500 Da
2.	Chemical ionization (CI), field ionization (FI)	Adduct ions: $[M + H]^+$	Non-polar, volatile, and some moderately polar molecules	~ 500 Da
3.	Fast atom bombardment (FAB)	Adduct ions: $[M + H]^+$ and $[M + Na]^+$	Polar, non-polar, volatile, and non-volatile molecules, e.g., peptides	~ 6000 Da
4.	Matrix-assisted laser desorption/ionization ^b (MALDI)	Adduct ions: $[M + H]^+$ and $[M + Na]^+$ “singly charged”	Mainly polar, non-volatile molecules, e.g., proteins and peptides	Up to ~100/150 kDa
5.	Electrospray ionization ^b (ESI)	Adduct ions: singly and multiply charged ions $[M + nH]^{n+}$, $n \geq 1$	Mainly polar, non-volatile molecules, e.g., proteins and peptides	Up to ~100/150 kDa

^aBased on [67, 74]^bMainly for proteins and peptides [41]. Also useful for polar lipids and carbohydrates

approach, which eventually aided immensely to be applied for a variety of biomolecules, including peptides [10].

Mass Analyzers

Besides ionization modes, development of mass analyzers of different configurations also had a major contribution for the significant advancement of the field of mass spectrometry, eventually facilitating high-throughput analysis and thereby paving way for the emergence of the different **omics** fields of study. Quadrupole (**Q**), time-of-flight (**ToF**), and ion trap (**IT**) are some examples of widely used mass analyzers [41, 67]. A simple mass spectrometer would consist of one type of ionization source coupled to a mass analyzer, which can be useful to measure m/z values of the ions that are generated in the ionization source. However, the use of mass analyzers is not only restricted to measure m/z values. It is possible to design and carry out certain specialized experiments in gas phase, involving different types of ions, when more than one mass analyzer is utilized in tandem with another. For example, it is possible to perform “tandem MS (MS/MS)” experiments, by using two or more mass analyzers. In an MS/MS experiment, population of a particular molecular ion (intact molecular ion) is selected (referred as “precursor ion”) and subjected to

fragmentation, resulting in the formation of fragment ions, whose m/z values are measured. The graphical plot of m/z values of the fragment ions and their respective abundances is called MS/MS spectrum of a particular precursor ion. The selected precursor ion would undergo fragmentation in multiple ways, depending on its molecular structure, viz., the nature of bonds and atoms constituting the molecule. The m/z values of the fragment ions would be useful for elucidation of molecular structure of unknown compounds. Hence, MS/MS has an important role for sequence elucidation of peptides.

As mentioned previously, in order to perform MS/MS experiments, two or more mass analyzers are required, for instance, triple quadrupole, which is an assembly of three quadrupoles. Other examples are Q-ToF and ToF-ToF, in which the second ToF is used for mass analysis of fragment ions that arise from a selected precursor ion. Thus, MALDI ToF-ToF and ESI Q-ToF are tandem mass spectrometers, in which MS/MS experiments can be carried out [41, 67]. However, it is possible to accomplish MS/MS experiments using one mass analyzer also, e.g., “three-electrode system” ion trap [61, 86, 87]. The key steps involved in the process of MS/MS experiments and data acquisition are schematically illustrated in Fig. 15.3. Since LC interfacing was compatible with both ESI and MALDI, it was feasible to perform LC-ESI-MS/MS and LC-MALDI-MS/MS experiments as well, which have been successfully applied to a variety of research studies [31, 75, 79, 88–90]. Consequently, the scope of high-throughput applications of MS enhanced significantly.

15.2.2.2 Applications of MS/MS and Databases

Among different methods of MS/MS known [91], collision induced dissociation (CID) has found widespread applications and has proven to be successful for deducing sequences of peptides from diverse biological sources [31, 46, 84, 89]. CID MS/MS studies conducted on several linear peptides of known sequences under the conditions of ESI and MALDI have shown that protonated precursor ions predominantly yield b- and y- type ions, whose structures are depicted in Fig. 15.4a [41, 92, 93]. Based on these structures, the equations to calculate the m/z values of b- and y- ions are shown in Fig. 15.4b. Thus, the sequences of linear peptides are usually derived using the m/z values of the detected b- and y- type ions in the MS/MS spectrum. However, it is also important to know that the fragmentation patterns of cyclic peptides are drastically different from that of linear peptides. Therefore, the equations to calculate the m/z values of fragment ions arising from cyclic peptide precursor ions would be different from those shown in Fig. 15.4b. In the context of “omics” studies, which involve high-throughput analysis, manual interpretation of CID MS/MS spectra of numerous peptides would be tedious and cumbersome. Consequently, computational approaches are sought after to simplify the process of sequence elucidation of peptides in omics investigations. Therefore, computational methods are employed to build databases in such a manner that it can interrogate with the several experimentally recorded MS/MS data of peptides and thereby their sequences are identified. Different types of databases have been constructed for this purpose. One way of constructing a database is to collate

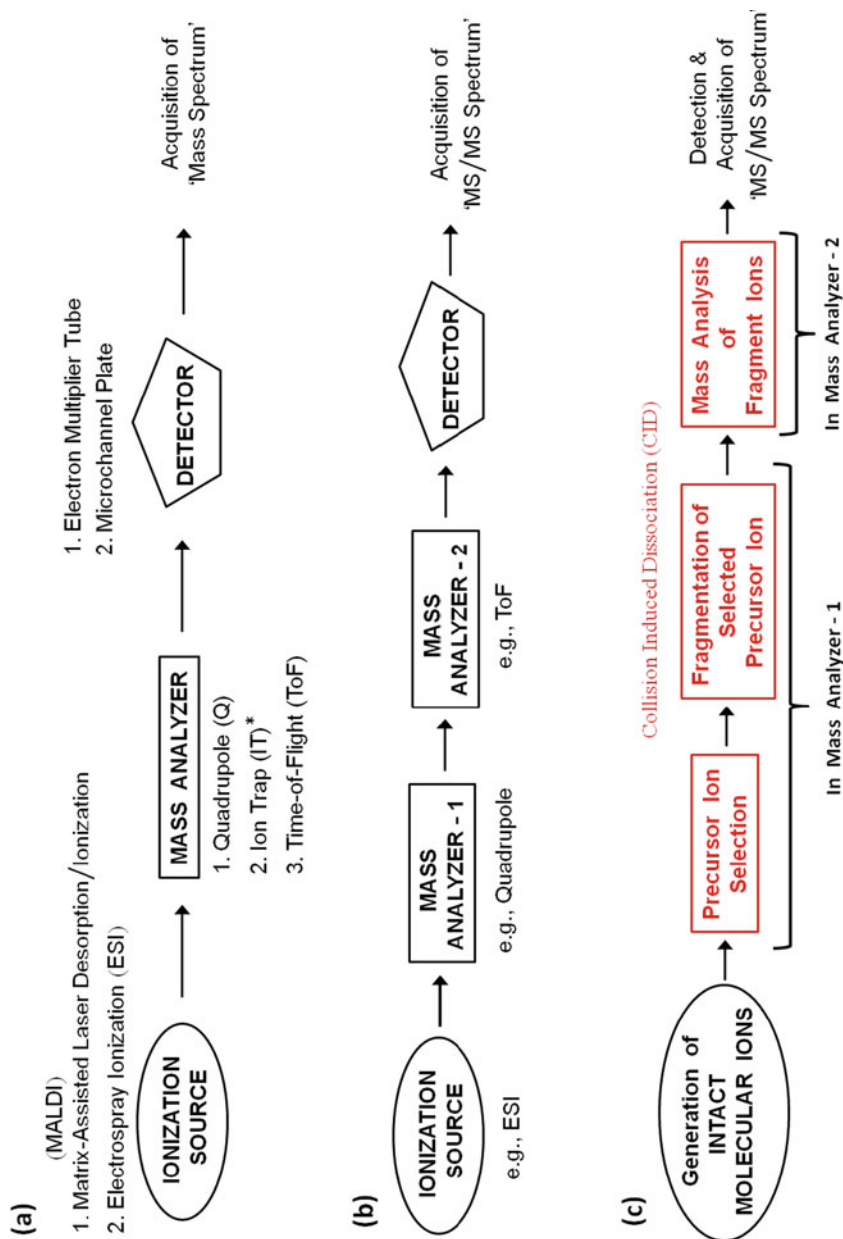


Fig. 15.3 Schematic illustration of major components of (a) a conventional mass spectrometer and (b) a tandem mass spectrometer. (c) Cartoon representation of key steps involved during acquisition of MS/MS spectrum in a tandem mass spectrometer

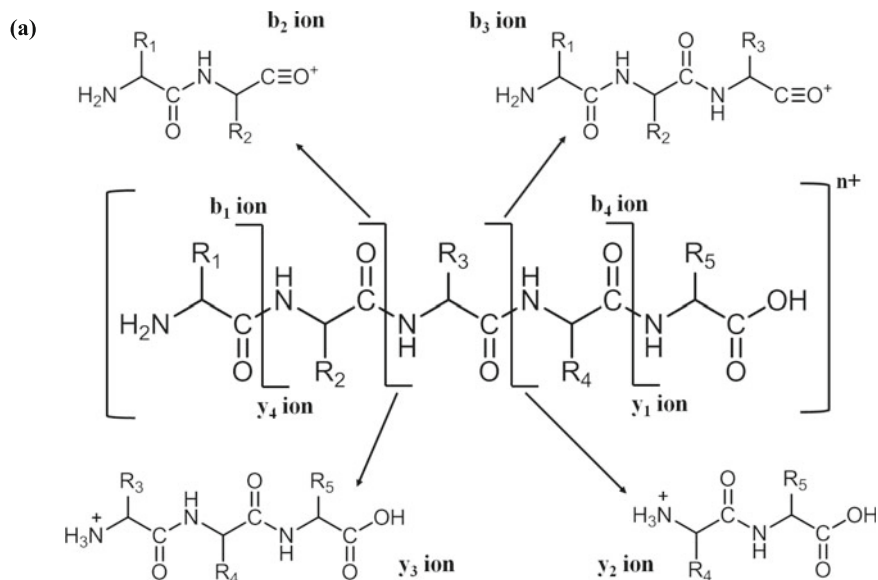


Fig. 15.4 (a) Depiction of backbone fragmentation of a linear pentapeptide precursor ion due to CID MS/MS, which predominantly gives rise to b- and y- type ions. Structures of b₂, b₃, y₂, and y₃ ions are shown. $n+$ refers to multiply protonated ion, where $n \geq 1$. (b) Equations to calculate the m/z values of singly charged b- and y- ions that result from CID MS/MS of linear peptide precursor ion. Monoisotopic masses rounded off to 4 places of decimal are shown. (c) Molecular structure of amino acid. Portion of the structure shown within the square bracket is called “amino acid residue” or “residue”. Residue mass = amino acid molecular mass – 18.0106 Daltons (monoisotopic mass)

experimental CID MS/MS spectra acquired on various peptides of “known” sequences; this can also be referred as building a library or spectral library [94, 95].

Another way of making a database involves theoretical calculation of m/z values of fragment ions (e.g., Fig. 15.4b), corresponding to various known peptide sequences, for which molecular masses of constituting amino acids are utilized in an appropriate manner. Alternatively, instead of known sequences, all “theoretically permuted peptide sequences” can also be considered to build a database that would consist of calculated fragment ions’ (b- and y- type ions) m/z values of the respective peptides. One more strategy to create a database is to utilize the experimentally determined transcriptome or genome sequences of a particular organism, from which peptide sequences can be predicted. Such predicted sequences can be used to generate m/z values of the peptide fragment ions, which then can be used to interrogate with the experimental CID MS/MS data for identifying peptides [90, 96–99]. Furthermore, even in the absence of any genome or transcriptome data or databases, peptides can be sequenced by “de novo” approach, which mostly involves manual interpretations of experimental MS/MS data, by using the molecular structure and masses of amino acids (Fig. 15.4). De novo approach can also involve some computational methods, which can aid in accelerating the process of elucidating the sequences [100–104].

15.3 Databases of Bioactive Peptides

A large compendium of literature is available on the sequence and characterization of bioactive peptides isolated from the invisible microbes to mammals including human. In the recent times, efforts in organizing this information from the literature to a structured set of data on a computer making it accessible to the users resulted in a number of databases. For instance, the database “PepBank” is derived by text mining of MEDLINE abstracts and other public resources like UniProtKB along with manual curation [105]. On the other hand, huge experimental data also have been arranged to construct databases. PeptideAtlas is an example of a database that is built using peptide sequences obtained from high-throughput mass spectrometric experiments carried out on eukaryotic organisms such as human, yeast, mouse, *Drosophila*, horse, cow, etc. [106]. However, this is more appropriate for proteomics, since all the peptides in this database are derived from proteins through proteolysis.

Among many classes of bioactive peptides, antimicrobial peptides (AMPs) constitute a major class. These antimicrobial peptides can further be grouped into the following major categories based on (1) specific sources, (2) targets (e.g., microbes), (3) selected human diseases, and (4) biosynthetic pathways, e.g., non-ribosomal route. Some of the well-known databases that have catalogued therapeutic peptides belonging to the above four categories are listed in Table 15.3. A general collection of AMPs from diverse sources are available at DBAASP [107], CAMP [108], and LAMP [109] databases, as they provide clues to understand the naturally occurring AMPs. A more comprehensive survey on the database of AMPs is discussed in

Table 15.3 A representative list of databases corresponding to different therapeutic peptide classes

Class of peptides	Source	Biological activity	Database (status)
Penaeidin	Shrimp	Antibacterial, antifungal	PenBase (inactive) http://www.penbase.immunaqua.com
Bacteriocins	Bacteria (gram-positive and gram-negative)	Antimicrobial	BACTIBASE (active) http://bactibase.hammamilab.org/
Plant AMPs	Plants	Antimicrobial	PhytAMP (active) http://phytamp.hammamilab.org/
Plant peptides	Plants	Antimicrobial, opioid, inhibitory, plant defense, etc.	PlantPepDB (active) http://www.nipgr.ac.in/PlantPepDB/
Peptaibols	Fungi (<i>Trichoderma</i> and <i>Emericelopsis</i>)	Antimicrobial	Peptaibol database (active) http://peptaibol.cryst.bbk.ac.uk/home.shtml
Conopeptides	Marine cone snails	Target ion channels of different subtypes	ConoServer (active) http://www.conoserver.org/
Cyclotides	Plants	Antimicrobial, insecticidal, anti-HIV	CyBase (active) http://www.cybase.org.au/
Inhibitor cystine knots	Plants, toxins from spider, cone snails, bug, crab, scorpion, and human	Antimicrobial, insecticidal, antitumor, antimalarial, antihelminthics, neurotransmitters	KNOTTIN (active) https://www.dsimb.inserm.fr/KNOTTIN/index.php
Non-ribosomal peptides	Bacteria, fungi, and marine sponge	Antimicrobial, surfactants, siderophore, etc.	NORINE (active) https://bioinfo.lifl.fr/norine/index.jsp
AMPs	Various sources	Antimicrobial	DBAASP (active) https://dbaasp.org/

(continued)

Table 15.3 (continued)

Class of peptides	Source	Biological activity	Database (status)
			CAMP (active) http://www.camp.bicnirrh.res.in/ LAMP (active) http://biotechlab.fudan.edu.cn/database/lamp/
Anticancer peptides	Various sources	Anticancer	CancerPPD (active) http://crdd.osdd.net/raghava/cancerppd/index.php
Antidiabetic peptides	Various sources	Antidiabetic	BioDADPep (active) http://omicsbase.com/BioDADPep
Tumor homing peptides	Various sources	Deliver drugs selectively in tumors	TumorHoPe (active) https://webs.iiitd.edu.in/raghava/tumorhope/
Cell penetrating peptides	Various sources	Delivery of biological cargos such as nucleic acids, proteins, etc.	CPPsite2.0 (active) https://webs.iiitd.edu.in/raghava/cppsitere
FDA approved/investigational peptides	Various sources	Peptide drugs—therapeutic effect	THPdb (active) http://crdd.osdd.net/raghava/thpdb/
Hemolytic and non-hemolytic peptides	Various sources	Antimicrobial, anticancer, antiviral, cell penetrating activity, etc.	Hemolytik (active) http://crdd.osdd.net/raghava/hemolytik/

[110]. Databases like PhytAMP [111], BACTIBASE [112], PenBase [113], and ConoServer [114, 115] have a collection of peptides produced by plants, bacteria, shrimp, and cone snails, respectively. Hence, these are a few examples of databases belonging to the first category of bioactive peptides.

Databases like ParaPep [116], AVPdb [117], AntiTbPdb [118], and BioDADPep [119] have been compiled using experimentally validated peptides targeting parasitic organisms, viruses, mycobacterium, and anti-diabetic peptides, respectively. CancerPPD [120], CPPsite 2.0 [121], and TumorHoPe [122] are a few databases, which have the collection of peptides applied for selective diseases, e.g., cancer. Some well-known AMPs that also have anticancer activity are magainins, cecropins, defensins, and pleurocidin [4]. The peptaibol database and NORINE are perhaps the only databases that have a comprehensive list of the peptides derived through the non-ribosomal routes [123, 124]. The peptaibol database contains peptides that are predominantly of fungal origin, while non-ribosomal peptides found in bacteria as well as fungi are catalogued in NORINE database. KNOTTIN is a unique database which provides standardized information on a specific family of peptides known as inhibitor cystine knots (ICKs) that possess potential therapeutic applications [125]. ICKs are structurally intriguing molecules, as they contain ultra-stable knot-like structures due to interwoven disulfide bridges. CyBase is another interesting database pertaining to plant cyclotides that are disulfide-bonded as well as backbone-cyclized RiPPs, which include ICKs too (see Fig. 15.1) [20]. The non-hemolytic activity is an important qualifier for a peptide to become a drug candidate. Therefore, the bioactive peptides listed in the above databases irrespective of their therapeutic potential could be classified further into two categories as follows: hemolytic and non-hemolytic peptides. Thus, a manually curated database of experimentally determined hemolytic and non-hemolytic peptides is already compiled in the Hemolytik database [126]. In addition to collection and organization of information, these databases also offer options for the users to perform a variety of “searches” so as to retrieve relevant information sought by them.

15.3.1 Web-Based Prediction Tools for Therapeutic Peptides

Unlike the databases which collect information on the bioactive peptides by data mining in the literature and many other sources, the prediction servers are developed based on thorough investigations on the sequence properties of these bioactive peptides. These prediction servers not only screen the peptides to study their sequence/structural features, but also provide directions to “design” bioactive peptides with improved therapeutic potential and reduced toxicity. A representative list of web-based prediction servers for therapeutic peptides is shown in Table 15.4. For example, prediction servers HemoPi [127] and HemoPred [128] are developed based on the Hemolytik database.

HemoPred is a web-based prediction server primarily based on the machine learning algorithms, wherein the sequence features of the peptides such as amino acid composition (AAC), dipeptide composition (DPC), and physicochemical properties (PCP) are assumed to be the primary descriptors that encode hemolytic peptides. On the other hand, HemoPi is developed to predict the hemolytic potency of peptides based on the study of the sequence features such as AAC, DPC, and “specific motifs” that are abundant in hemolytic peptides. HemoPi is available as a

Table 15.4 A representative list of web-based prediction servers for therapeutic peptides

Class of peptides	Source	Biological activity	Prediction server (Status)
Antiviral peptides	Various sources	Antiviral	AVPpred (inactive) http://crdd.osdd.net/servers/avppred
CyPred	Archaea, Bacteria, and Eukaryotic proteomes	Antimicrobial	CyPred (active) http://biomine.cs.vcu.edu/servers/CyPred/
RiPPs	Various sources	Antimicrobial	RiPPMiner (active) http://www.nii.ac.in/~priyesh/lantipepDB/new_predictions/index.php
Hemolytic and non-hemolytic peptides	Various sources	Antimicrobial, anticancer, antiviral, cell penetrating activity, etc.	HemoPi (inactive) HemoPred (active) http://codes.bio/hemopred/ HAPPENN (active) https://research.timmons.eu/happenn HLPpred-Fuse (active) http://thegleelab.org/HLPpred-Fuse/index.html

web server, mobile app, as well as a JAVA-based standalone software. Other web servers for the prediction of hemolytic and non-hemolytic peptides and the level of hemolytic activity (high/low) include HAPPENN [129] and HLPpred-Fuse [130]. CyPred [131] is a web server available for the prediction of cyclic peptides based on sequences. In addition to a web-based predictor, this server also has a database of putative cyclic peptides predicted from Archaea, Bacteria, and Eukaryotic proteomes. RiPPMiner [132] is a web-based tool developed to decode the chemical structures of RiPPs. The predictive power of RiPPMiner is based on the training set containing manually curated sequences (~ 500 sequences) belonging to 13 subclasses of peptides. Thus, CyPred and RiPPMiner could be considered as structure prediction servers.

15.3.2 Databases/Prediction Tools Only Narrow the Search Space

Databases/libraries are rather useful for rapid screening and filtering purposes, thereby aiding in minimizing the time of manual data analysis. Therefore, it should not be reckoned that use of databases can substitute manual interpretation. In other words, manual interpretations cannot be completely discounted for identifying peptides using MS/MS, regardless of the usage of any type of database. However, manual interpretation requires a suitable level of expertise. Thus, in several cases manual analysis cannot be evaded. Also, for many cases, both manual- and database-based automated approach are required to an equal extent so as to minimize the

errors during analysis. There have been (can be) instances when elucidation of novel sequences requires more involvement of manual interpretations, especially when it is not possible to match a particular experimental MS/MS data with any database. This means, not finding suitable entry in any type of database for experimental data need not be considered as discouraging clue, as this might also imply that those experimental data are perhaps alluding to a novel molecule, which may not be available in any already existing databases. However, it is also very important to ensure that such experimental data are reproducible and of good quality, acquired from well-prepared sample.

15.4 Future Prospects

The hunt for peptide-based drug compounds has not ceased, for which diverse biological sources are continuously being explored time and again. This is primarily because of growing incidence of antimicrobial and multidrug resistance with the conventional antibiotics [133, 134]. Since less resistance is often encountered with peptides than with traditional antibiotics, peptidic drugs, especially antimicrobial peptides, are regarded as novel alternative agents [47, 135]. The antimicrobial peptides elicit function similar to or even better than the usual antibiotics [136]. Moreover, several peptides have been observed to possess broad-spectrum antimicrobial specificity [134, 135]. However, one of the major drawbacks with peptides is that they are highly susceptible to degradation *in vivo*, due to proteolysis and hence, have very short elimination half-lives in plasma, due to which they eventually lose their potency [133, 137]. Consequently, many a times, several peptidic compounds fail or are suspended, during the preclinical or clinical trials, although these compounds might exhibit very good activity in various kinds of *in vitro* studies [133, 134]. Different strategies have been attempted for resolving the issue of instability toward proteolysis by suitably modifying certain selected sites on the peptide structure, which might render them resistant against proteolysis [135, 137–139]. Such modified compounds (semi-synthetic analogues of fully synthetic compounds) could therefore elicit better activity and eventually could become successful drugs in the market. Also, various kinds of carriers have been developed, particularly employing nanotechnological methods, in order to protect the peptide-based bioactive compounds from degradation and deliver at the required site in the body [134].

Due to the urgent requirement of novel peptide-based antimicrobial compounds, there is a growing demand for the “methods” that can “swiftly” identify peptide molecules. This suggests that the need for “omics” approach is inevitable, as it encompasses methods that are particularly tailored for high-throughput analysis and thereby accelerate the process of peptide drug discovery. Since chromatography and mass spectrometry allow high-throughput screening of peptides directly from the crude biological extracts at a faster pace, these two techniques have become integral for peptidomics-based drug discovery. Another important requirement in peptidomics research is *database*, which enables rapid analysis of the experimental

data. In fact, databases are required for other branches of omics studies as well, e.g., genomics and transcriptomics. Further, in the current scenario, X-ray crystallography and NMR spectroscopy do not exactly come under omics science; however, these techniques might be adapted so as to make them suitable for omics approach in the near future.

References

1. Berditsch M, Lux H, Babii O, Afonin S, Ulrich AS (2016) Therapeutic potential of gramicidin S in the treatment of root canal infections. *Pharmaceuticals* 9:56. <https://doi.org/10.3390/ph9030056>
2. de la Torre BG, Albericio F (2020) Peptide therapeutics 2.0. *Molecules* 25(10):2293
3. McGivern JG (2007) Ziconotide: a review of its pharmacology and use in the treatment of pain. *Neuropsychiatr Dis Treat* 3(1):69–85
4. Boohaker RJ, Lee MW, Vishnubhotla P, Perez JM, Khaled AR (2012) The use of therapeutic peptides to target and to kill Cancer cells. *Curr Med Chem* 19(22):3794–3804
5. Kim IW, Kim SJ, Kwon YN, Yun EY, Ahn MY, Kang DC, Hwang JS (2012) Effects of the synthetic coprisin analog peptide, CopA3 in pathogenic microorganisms and mammalian cancer cells. *J Microbiol Biotechnol* 22(1):156–158. <https://doi.org/10.4014/jmb.1109.09014>
6. Lefèbvre PJ (2003) Biosynthesis, secretion, and action of glucagon. In international textbook of diabetes mellitus (eds DeFronzo, Ferraninni, keen and Zimmet). <https://doi.org/10.1002/0470862092.d0208>
7. Zhangsun D, Luo S, Wu Y, Zhu X, Hu Y, Xie L (2006) Novel O-superfamily Conotoxins identified by cDNA cloning from three Vermivorous Conus species. *Chem Biol Drug Des* 68:256–265
8. Dang T, Süßmuth RD (2017) Bioactive peptide natural products as Lead structures for medicinal use. *Acc Chem Res* 50:1566–1576
9. Gimenez-Gallego G, Navia MA, Reuben JP, Katz GM, Kaczorowski GJ, Garcia ML (1988) Purification, sequence, and model structure of charybdotoxin, a potent selective inhibitor of calcium-activated potassium channels. *Proc Natl Acad Sci U S A* 85(10):3329–3333
10. Li L, Sweedler JV (2008) Peptides in the brain: mass spectrometry-based measurement approaches and challenges. *Annu Rev Anal Chem* 1:451–483
11. Terlau H, Olivera BM (2004) Conus venoms: a rich source of novel ion channel-targeted peptides. *Physiol Rev* 84(1):41–68. <https://doi.org/10.1152/physrev.00020.2003>
12. Kleinkauf H, Von Döhren H (1996) A nonribosomal system of peptide biosynthesis. *Eur J Biochem* 236:335–351. <https://doi.org/10.1111/j.1432-1033.1996.00335.x>
13. Marahiel MA, Stachelhaus T, Mootz HD (1997) Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem Rev* 97:2651–2673. <https://doi.org/10.1021/cr960029e>
14. Krause C, Kirschbaum J, Brückner H (2006) Peptaibiotics: an advanced, rapid and selective analysis of peptaibiotics/peptaibols by SPE/LC-ES-MS. *Amino Acids* 30:435–443. <https://doi.org/10.1007/s00726-005-0275-9>
15. Prasad BVV, Balaran P (1984) The stereochemistry of peptides containing α -aminoisobutyric acid. *CRC Crit Rev Biochem* 16(4):307–348
16. Vizán JL, Hernández-Chico C, del Castillo I, Moreno F (1991) The peptide antibiotic microcin B17 induces double-strand cleavage of DNA mediated by *E. coli* DNA gyrase. *EMBO J* 10(2):467–476
17. Yorgey P, Lee J, Kördel J, Vivas E, Warner P, Jebaratnam D, Kolter R (1994) Posttranslational modifications in microcin B17 define an additional class of DNA gyrase inhibitor. *Proc Natl Acad Sci U S A* 91(10):4519–4523. <https://doi.org/10.1073/pnas.91.10.4519>
18. Matsumoto T, Shishido A, Morita H, Itokawa H, Takeya K (2001) Cyclolinopeptides F–I, cyclic peptides from linseed. *Phytochemistry* 57:251–260

19. Safavi-Hemami H, Brogan SE, Olivera BM (2019) Pain therapeutics from cone snail venoms: from Ziconotide to novel non-opioid pathways. *J Proteomics* 190:12–20. <https://doi.org/10.1016/j.jprot.2018.05.009>
20. Wang CK, Kaas Q, Chiche L, Craik DJ (2008) CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Res* 36 (Database issue):D206–D210. <https://doi.org/10.1093/nar/gkm953>
21. Mulvenna JP, Wang C, Craik DJ (2006) CyBase: a database of cyclic protein sequence and structure. *Nucleic Acids Res* 34(Database issue):D192–D194. <https://doi.org/10.1093/nar/gkj005>
22. Hegemann JD, Zimmermann M, Xie X, Marahiel MA (2015) Lasso peptides: an intriguing class of bacterial natural products. *Acc Chem Res* 48(7):1909–1919. <https://doi.org/10.1021/acs.accounts.5b00156>
23. Maksimov MO, Pan SJ, Link AJ (2012) Lasso peptides: structure, function, biosynthesis, and engineering. *Nat Prod Rep* 29(9):996–1006. <https://doi.org/10.1039/c2np20070h>
24. Fox RO Jr, Richards FM (1982) A voltage-gated ion channel model inferred from the crystal structure of alamethicin at 1.5-Å resolution. *Nature* 300(5890):325–330. <https://doi.org/10.1038/300325a0>
25. Mathew MK, Balam P (1983) Alamethicin and related membrane channel forming polypeptides. *Mol Cell Biochem* 50(1):47–64. <https://doi.org/10.1007/BF00225279>
26. Kessler N, Schuhmann H, Morneweg S, Linne U, Marahiel MA (2004) The linear pentadecapeptide gramicidin is assembled by four multimodular nonribosomal peptide synthetases that comprise 16 modules with 56 catalytic domains. *J Biol Chem* 279(9):7413–7419. <https://doi.org/10.1074/jbc.M309658200>
27. David JM, Rajasekaran AK (2015) Gramicidin A: a new Mission for an old antibiotic. *J Kidney Cancer VHL* 2(1):15–24. <https://doi.org/10.15586/jkcvhl.2015.21>
28. Guan Q, Huang S, Jin Y, Campagne R, Alezra V, Wan Y (2019) Recent advances in the exploration of therapeutic analogues of gramicidin S, an old but still potent antimicrobial peptide. *J Med Chem* 62(17):7603–7617. <https://doi.org/10.1021/acs.jmedchem.9b00156>
29. Kobayashi J, Nakamura T, Tsuda M (1996) Hymenamamide F, new cyclic heptapeptide from marine sponge Hymeniacidon sp. *Tetrahedron* 52(18):6355–6360. [https://doi.org/10.1016/0040-4020\(96\)00281-5](https://doi.org/10.1016/0040-4020(96)00281-5)
30. Arima K, Kakinuma A, Tamura G (1968) Surfactin, a crystalline peptide lipid surfactant produced by *Bacillus subtilis*: isolation, characterization and its inhibition of fibrin clot formation. *Biochem Biophys Res Commun* 31(3):488–494. [https://doi.org/10.1016/0006-291x\(68\)90503-2](https://doi.org/10.1016/0006-291x(68)90503-2)
31. Pecci Y, Rivardo F, Martinotti MG, Allegrone G (2010) LC/ESI-MS/MS characterization of lipopeptide biosurfactants produced by *Bacillus licheniformis* V9T14 strain. *J Mass Spectrom* 45:772–778
32. Bertero A, Fossati P, Tedesco DEA, Caloni F (2020) Beauvericin and Enniatins: *In Vitro* intestinal effects. *Toxins (Basel)* 12(11):686. <https://doi.org/10.3390/toxins12110686>
33. Pathak KV, Keharia H, Gupta K, Thakur SS, Balam P (2012) Lipopeptides from the banyan endophyte, *Bacillus subtilis* K1: mass spectrometric characterization of a library of fengycins. *J Am Soc Mass Spec* 23(10):1716–1728. <https://doi.org/10.1021/jasms.8b04151>
34. Hamann MT, Scheuer PJ (1993) Kahalalide F: a bioactive depsipeptide from the sacoglossan mollusk *Elysia rufescens* and the green alga *Bryopsis* sp. *J Am Chem Soc* 115(13):5825–5826. <https://doi.org/10.1021/ja00066a061>
35. Aboye TL, Camarero JA (2012) Biological Synthesis of Circular Polypeptides. *J Biol Chem* 287(32):27026–27032
36. Amison PG, Bibb MJ, Bierbaum G et al (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* 30:108–160
37. Zhao BC, Lin HC, Yang D, Ye X, Li ZG (2015) Disulfide Bridges in Defensins. *Curr Top Med Chem* 16(2):206–19. <https://doi.org/10.2174/1568026615666150701115911>

38. Kleinkauf H, Von Döhren H (1987) Biosynthesis of peptide antibiotics. *Annu Rev Microbiol* 41(1):259–289
39. Simpson RJ (2003) *Proteins and Proteomics: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, New York. ISBN 0-87969-554-4
40. Schrader M, Schulz-Knappe P (2001) Peptidomics technologies for human body fluids. *Trends Biotechnol* 19:S55–S60
41. Kinter M, Sherman NE (2000) *Protein Sequencing and Identification using Tandem Mass Spectrometry*. John Wiley & Sons, New York
42. Bennet HPJ, Browne CA, Solomon S (1981) Purification of the two major forms of rat pituitary Corticotropin using only reversed-phase liquid chromatography. *Biochemistry* 20:4530–4538
43. Xi X, Li R, Jiang Y, Lin Y, Wu Y, Zhou M, Xu J, Wang L, Chen T, Shaw C (2013) Medusins: a new class of antimicrobial peptides from the skin secretions of phyllomedusine frogs. *Biochimie* 95(6):1288–1296
44. Ye X, Zhao N, Yu X, Han X, Gao H, Zhang X (2016) Extensive characterization of peptides from *Panax ginseng* C. A. Meyer using mass spectrometric approach. *Proteomics* 16 (21):2788–2791
45. Bennet HPJ (1983) Isolation of pituitary peptides by reversed-phase high-performance liquid chromatography: expansion of the resolving power of reversed-phase columns by manipulating pH and the nature of the ion-pairing reagent. *J Chromatogr A* 266:501–510. [https://doi.org/10.1016/S0021-9673\(01\)90921-5](https://doi.org/10.1016/S0021-9673(01)90921-5)
46. Yang X, Huang E, Yuan C, Zhang L, Yousef AM (2016) Isolation and structural elucidation of Brevibacillin, an antimicrobial Lipopeptide from *Brevibacillus laterosporus* that combats drug-resistant gram-positive Bacteria. *Appl Environ Microbiol* 82(9):2763–2772
47. Hansen IKØ, Isaksson J, Poth AG, Hansen KØ, Andersen AJC, Richard CSM, Blencke H-M, Stensvåg K, Craik DJ, Haug T (2020) Isolation and characterization of antimicrobial peptides with unusual disulfide connectivity from the colonial ascidian *Synoicum turgens*. *Mar Drugs* 18:51. <https://doi.org/10.3390/md18010051>
48. Smith C, Phillips M, Miller C (1986) Purification of Charybdotoxin, a specific inhibitor of the high conductance Ca^{2+} –activated K^{+} channel. *J Biol Chem* 261(31):14607–14613
49. Conlon JM (2007) Purification of naturally occurring peptides by reversed-phase HPLC. *Nat Protoc* 2(1):191–197
50. Mant CT, Chen Y, Yan Z, Popa TV, Kovacs JM, Mills JB, Tripet BP, Hodges RS (2007) HPLC analysis and purification of peptides. In: *Peptide characterization and application protocols* (Ed. Fields G), *Methods in Molecular Biology*, vol 386. Humana Press Inc., Totowa, NJ
51. Brückner H, Przybylski M (1984) Isolation and structural characterization of polypeptide-antibiotics of the peptaibol class by high-performance liquid chromatography with field desorption and fast atom bombardment mass spectrometry. *J Chromatogr* 296:263–275
52. Chua VM, Gajewiak J, Watkins M, Espino SS, Ramiro IBL, Omega CA, Imperial JS, Carpio LPD, Fedosov A, Safavi-Hemami H, Salvador-Reyes LA, Olivera BM, Concepcion GP (2020) Purification and characterization of the pink-Floyd Drillipeptide, a bioactive venom peptide from *Clavus davidgilmouri* (Gastropoda: Conoidea: Drilliliidae). *Toxins (Basel)* 12 (8):508
53. Neves JLB, Imperial JS, Morgenstern D, Ueberheide B, Gajewiak J, Antunes A, Robinson SD, Espino S, Watkins M, Vasconcelos V, Olivera BM (2019) Characterization of the first Conotoxin from *Conus ateralbus*, a Vermivorous cone snail from the Cabo Verde archipelago. *Mar Drugs* 17(8):432
54. Chicchi GG, Gimenez-Gallego G, Ber E, Garcia ML, Winquist R, Cascieri MA (1988) Purification and characterization of a unique, potent inhibitor of Apamin binding from *Leiurus quinquestriatus hebraeus* venom. *J Biol Chem* 263(21):10192–10197

55. Guo D, Mant CT, Hodges RS (1987) Effects of ion-pairing reagents on the prediction of peptide retention in reversed-phase high-resolution liquid chromatography. *J Chromatogr A* 386:205–222. [https://doi.org/10.1016/S0021-9673\(01\)94598-4](https://doi.org/10.1016/S0021-9673(01)94598-4)
56. Emwas A-H, Roy R, McKay RT, Tenori L, Saccenti E, Gowda GAN, Raftery D, Alahmari F, Jaremkó L, Jaremkó M, Wishart DS (2019) NMR spectroscopy for metabolomics research. *Metabolites* 9(7):123. <https://doi.org/10.3390/metabo9070123>
57. Mahrous EA, Farag MA (2015) Two dimensional NMR spectroscopic approaches for exploring plant metabolome: A review. *J Adv Res* 6(1):3–15
58. Tomou E-M, Chatziathanasiadou MV, Chatzopoulou P, Tzakos AG, Skaltsa H (2020) NMR-based chemical profiling, isolation and evaluation of the cytotoxic potential of the Diterpenoid Siderol from cultivated *Sideritis euboea* Heldr. *Molecules* 25:2382. <https://doi.org/10.3390/molecules25102382>
59. Spencer RK, Nowick JS (2015) A Newcomer's guide to peptide crystallography. *Isr J Chem* 55(6–7):698–710
60. Peng J, Gygi SP (2001) Proteomics: the move to mixtures. *J Mass Spectrom* 36(10):1083–1091
61. Jonscher KR, Yates JR III (1997) The quadrupole ion trap mass spectrometer--a small solution to a big challenge. *Anal Biochem* 244(1):1–15
62. Pandey RC, Cook JC, Rinehart KL Jr (1977) High resolution and field desorption mass spectrometry studies and revised structures of alamethicins I and II. *J Am Chem Soc* 99:8469–8483
63. Pandey RC, Cook JC Jr, Rinehart KL Jr (1977) Reptaibophol antibiotics. 2. Structures of the peptide antibiotics emerimicins III and IV. *J Am Chem Soc* 99:5205–5206
64. Pandey RC, Meng H, Cook JC Jr, Rinehart KL Jr (1977) Structure of antiamoebin I from high resolution field desorption and gas chromatographic mass spectrometry studies. *J Am Chem Soc* 99:5203–5205
65. Barber M, Bordoli RS, Sedgwick RD, Tyler AN (1981) Fast atom bombardment of solids as an ion source in mass spectrometry. *Nature* 293:270–275
66. Barber M, Bordoli RS, Sedgwick RD, Tyler AN (1981) Fast atom bombardment of solids (F.A.B.): a new ion source for mass spectrometry. *J Chem Soc Chem Commun* 325–327
67. de Hoffmann E, Stroobant V (2007) *Mass Spectrometry: Principles and Applications*, 3rd edn. John Wiley & Sons, Ltd
68. Leclerc G, Goulard C, Prigent Y, Bodo B, Wróblewski HRS (2001) Sequences and Antimycoplasmic properties of Longibrachins LGB II and LGB III, two novel 20-residue Peptaibols from *Trichoderma longibrachiatum*. *J Nat Prod* 64:164–170
69. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246:64–71
70. Karas M, Hillenkamp E (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* 60:2299–2301
71. Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T (1988) *Rapid Commun Mass Spectrom* 2:151–153
72. Fenn JB (2003) Electrospray wings for molecular elephants (Nobel Lecture). *Angew Chem Int Ed Engl* 42(33):3871–3894
73. Tanaka K (2003) The origin of macromolecule ionization by laser irradiation (Nobel Lecture). *Angew Chem Int Ed Engl* 42(33):3860–3870
74. Vestal ML (2001) Methods of ion generation. *Chem Rev* 101:361–375
75. Ducret A, Oostveen IV, Eng JK, Yates JR 3rd, Aebersold R (1998) High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. *Protein Sci* 7(3):706–719
76. Whitehouse CM, Dreyer RN, Yamashita M, Fenn JB (1985) Electrospray interface for liquid chromatographs and mass spectrometers. *Anal Chem* 57:675–679
77. Mandal AK, Ramasamy MRS, Sabareesh V, Openshaw ME, Krishnan KS, Balaram P (2007) Sequencing of T-superfamily Conotoxins from *Conus virgo*: Pyroglutamic acid identification

- and disulfide arrangement by MALDI mass spectrometry. *J Am Soc Mass Spectrom* 18:1396–1404
78. Ravindra G, Ranganayaki RS, Raghothama S, Srinivasan MC, Gilardi RD, Karle IL, Balam P (2004) Two novel hexadepsipeptides containing several modified amino acid residues from the fungus *Isaria*. *Chem Biodivers* 1:489–504
79. Sabareesh V, Ranganayaki RS, Raghothama S, Bopanna MP, Balam H, Srinivasan MC, Balam P (2007) Identification and characterization of a library of microheterogenous cyclohexadepsipeptides from the fungus *Isaria*. *J Nat Prod* 70(5):715–729
80. Janssen KA, Coradin M, Lu C, Sidoli S, Garcia BA (2019) Quantitation of single and combinatorial histone modifications by integrated chromatography of bottom-up peptides and middle-down polypeptide tails. *J Am Soc Mass Spectrom* 30(12):2449–2459
81. Pandeswari PB, Sabareesh V (2019) Middle-down approach: a choice to sequence and characterize proteins/proteomes by mass spectrometry. *RSC Adv* 9:313–344
82. Bodnar WM, Blackburn RK, Krise JM, Moseley MA (2003) Exploiting the complementary nature of LC/MALDI/MS/MS and LC/ESI/MS/MS for increased proteome coverage. *J Am Soc Mass Spectrom* 14(9):971–979
83. Pereira F, Niu X, deMello AJ (2013) A Nano LC-MALDI mass spectrometry droplet Interface for the analysis of complex protein samples. *PLoS One* 8(5):e63087. <https://doi.org/10.1371/journal.pone.0063087>
84. Hofmann S, Gluckmann M, Kausche S, Schmidt A, Corvey C, Lichtenfels R, Huber C, Albrecht C, Karas M, Herr W (2005) Rapid and sensitive identification of major histocompatibility complex class I-associated tumor peptides by Nano-LC MALDI MS/MS. *Mol Cell Proteomics* 4:1888–1897
85. Hölttä M, Zetterberg H, Mirgorodskaya E, Mattsson N, Blennow K, Gobom J (2012) Peptidome analysis of cerebrospinal fluid by LC-MALDI MS. *PLoS One* 7(8):e42555
86. March RE (1997) An Introduction to quadrupole ion trap mass spectrometry. *J Mass Spectrom* 32(4):351–369
87. March RE (2000) Quadrupole ion trap mass spectrometry. In: *Encyclopedia of Analytical Chemistry* (Ed. Meyers RA), vol 113. John Wiley & Sons, pp 11848–11872
88. Jaworski A, Brückner H (1999) Detection of new sequences of peptaibol antibiotics trichothoxins A-40 by on-line liquid chromatography-electrospray ionization mass spectrometry. *J Chromatogr A* 862:179–189. [https://doi.org/10.1016/S0021-9673\(99\)00931-0](https://doi.org/10.1016/S0021-9673(99)00931-0)
89. Martinez AFC, Moraes LAB (2015) Liquid chromatography-tandem mass spectrometry characterization of five new leucinostatin produced by *Paecilomyces lilacinus* CG—189. *J Antibiot* 68:178–184
90. Vijayarathy M, Basheer SM, Franklin JB, Balam P (2017) Contryphan genes and mature peptides in the venom of nine cone snail species by transcriptomic and mass spectrometric analysis. *J Proteome Res* 16(2):763–772
91. Brodbelt JS (2016) Ion activation methods for peptides and proteins. *Anal Chem* 88(1):30–51
92. Biemann K (1990) Appendix 5. Nomenclature for Peptide Fragment Ions (Positive Ions). *Methods Enzymol* 193:886–887
93. Roepstorff P, Fohlman J (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* 11:601
94. Craig R, Cortens JC, Fenyo D, Beavis RC (2006) Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res* 5(8):1843–1849
95. Fälth M, Svensson M, Nilsson A, Sköld K, Fenyo D, Andren PE (2008) Validation of endogenous peptide identifications using a database of tandem mass spectra. *J Proteome Res* 7(7):3049–3053
96. Gimenez G, Metcalf P, Paterson NG, Sharpe ML (2016) Mass spectrometry analysis and transcriptome sequencing reveal glowing squid crystal proteins are in the same superfamily as firefly luciferase. *Sci Rep* 6:27638
97. Hayakawa E, Watanabe H, Menschaert G, Holstein TW, Baggerman G, Schoofs L (2019) A combined strategy of neuropeptide prediction and tandem mass spectrometry identifies

- evolutionarily conserved ancient neuropeptides in the sea anemone *Nematostella vectensis*. PLoS One 14(9):e0215185. <https://doi.org/10.1371/journal.pone.0215185>
98. Kalmankar NV, Venkatesan R, Balaram P, Sowdhamini R (2020) Transcriptomic profiling of the medicinal plant *Clitoria ternatea*: identification of potential genes in cyclotide biosynthesis. Sci Rep 10:12658
 99. Robinson SD, Undheim EAB, Ueberheide B, King GF (2017) Venom peptides as therapeutics: advances, challenges and the future of venom-peptide discovery. Expert Rev Proteomics 14(10):931–939
 100. Allmer J (2011) Algorithms for the de novo sequencing of peptides from tandem mass spectra. Expert Rev Proteomics 8(5):645–657
 101. Blank-Landeshammer B, Kollipara L, Biß K, Pfenninger M, Malchow S, Shuvaev K, Zahedi RP, Sickmann A (2017) Combining De novo peptide sequencing algorithms, a synergistic approach to boost both identifications and confidence in bottom-up proteomics. J Proteome Res 16(9):3209–3218. <https://doi.org/10.1021/acs.jproteome.7b00198>
 102. Jagannath S, Sabareesh V (2007) Peptide Fragment Ion Analyzer (PFIA): a simple and versatile tool for the interpretation of tandem mass spectrometric data and de novo sequencing of peptides. Rapid Commun Mass Spectrom 21:3033–3038
 103. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 17(20):2337–2342. <https://doi.org/10.1002/rcm.1196>
 104. Tran NH, Zhang X, Xin L, Shan B, Li M (2017) De novo peptide sequencing by deep learning. Proc Natl Acad Sci U S A 114(31):8247–8252
 105. Shtatland T, Guettler D, Kossodo M, Pivovarov M, Weissleder R (2007) PepBank - a database of peptides based on sequence text mining and public peptide data sources. BMC Bioinformatics 8:280. <https://doi.org/10.1186/1471-2105-8-280>
 106. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Edde J, Loevenich SN, Aebersold R (2006) The PeptideAtlas project. Nucleic Acids Res 34(Database issue):D655–D658. <https://doi.org/10.1093/nar/gkj040>
 107. Gogoladze G, Grigolava M, Vishnepolsky B, Chubinidze M, Duroux P, Lefranc MP, Pirtskhalava M (2014) DBAASP: database of antimicrobial activity and structure of peptides. FEMS Microbiol Lett 357(1):63–68. <https://doi.org/10.1111/1574-6968.12489>
 108. Waghlu FH, Gopi L, Barai RS, Ramteke P, Nizami B, Idicula-Thomas S (2014) CAMP: collection of sequences and structures of antimicrobial peptides. Nucleic Acids Res 42(Database issue):D1154–D1158. <https://doi.org/10.1093/nar/gkt1157>
 109. Zhao X, Wu H, Lu H, Li G, Huang Q (2013) LAMP: a database linking antimicrobial peptides. PLoS One 8(6):e66557. <https://doi.org/10.1371/journal.pone.0066557>
 110. Wang G (2020) The antimicrobial peptide database provides a platform for decoding the design principles of naturally occurring antimicrobial peptides. Protein Sci 29(1):8–18. <https://doi.org/10.1002/pro.3702>
 111. Hammami R, Ben Hamida J, Vergoten G, Fliss I (2009) PhytAMP: a database dedicated to antimicrobial plant peptides. Nucleic Acids Res 37(Database issue):D963–D968. <https://doi.org/10.1093/nar/gkn655>
 112. Hammami R, Zouhir A, Ben Hamida J, Fliss I (2007) BACTIBASE: a new web-accessible database for bacteriocin characterization. BMC Microbiol 7:89. <https://doi.org/10.1186/1471-2180-7-89>
 113. Gueguen Y, Garnier J, Robert L et al (2006) PenBase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. Dev Comp Immunol 30(3):283–288. <https://doi.org/10.1016/j.dci.2005.04.003>
 114. Kaas Q, Westermann JC, Halai R, Wang CK, Craik DJ (2008) ConoServer, a database for conopeptide sequences and structures. Bioinformatics 24(3):445–446. <https://doi.org/10.1093/bioinformatics/btm596>

115. Kaas Q, Yu R, Jin AH, Dutertre S, Craik DJ (2012) ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res* 40(Database issue):D325–D330. <https://doi.org/10.1093/nar/gkr886>
116. Mehta D, Anand P, Kumar V, Joshi A, Mathur D, Singh S, Tuknait A, Chaudhary K, Gautam SK, Gautam A, Varshney GC, Raghava GP (2014) ParaPep: a web resource for experimentally validated antiparasitic peptide sequences and their structures. *Database (Oxford)* 2014:bau051. <https://doi.org/10.1093/database/bau051>
117. Qureshi A, Thakur N, Tandon H, Kumar M (2014) AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res* 42(Database issue):D1147–D1153. <https://doi.org/10.1093/nar/gkt1191>
118. Usmani SS, Kumar R, Kumar V, Singh S, Raghava GPS (2018) AntiTbPdb: a knowledgebase of anti-tubercular peptides. *Database (Oxford)*. 2018:bay025. <https://doi.org/10.1093/database/bay025>
119. Roy S, Teron R (2019) BioDADPep: a bioinformatics database for anti-diabetic peptides. *Bioinformation* 15(11):780–783. <https://doi.org/10.6026/97320630015780>
120. Tyagi A, Tuknait A, Anand P, Gupta S, Sharma M, Mathur D, Joshi A, Singh S, Gautam A, Raghava GP (2015) CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res* 43(Database issue):D837–D843. <https://doi.org/10.1093/nar/gku892>
121. Agrawal P, Bhalla S, Usmani SS, Singh S, Chaudhary K, Raghava GP, Gautam A (2016) CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. *Nucleic Acids Res* 44(D1):D1098–D1103. <https://doi.org/10.1093/nar/gkv1266>
122. Kapoor P, Singh H, Gautam A, Chaudhary K, Kumar R, Raghava GP (2012) TumorHoPe: a database of tumor homing peptides. *PLoS One* 7(4):e35187. <https://doi.org/10.1371/journal.pone.0035187>
123. Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P, Kucherov G (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res* 36(Database issue):D326–D331. <https://doi.org/10.1093/nar/gkm792>
124. Whitmore L, Wallace BA (2004) The Peptaibol database: a database for sequences and structures of naturally occurring peptaibols. *Nucleic Acids Res* 32(Database issue):D593–D594
125. Gelly JC, Gracy J, Kaas Q, Le-Nguyen D, Heitz A, Chiche L (2004) The KNOTTIN website and database: a new information system dedicated to the knottin scaffold. *Nucleic Acids Res* 32(Database issue):D156–D159. <https://doi.org/10.1093/nar/gkh015>
126. Gautam A, Chaudhary K, Singh S, Joshi A, Anand P, Tuknait A, Mathur D, Varshney GC, Raghava GP (2014) Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Res* 42(Database issue):D444–D449. <https://doi.org/10.1093/nar/gkt1008>
127. Chaudhary K, Kumar R, Singh S, Tuknait A, Gautam A, Mathur D, Anand P, Varshney GC, Raghava GP (2016) A web server and Mobile app for computing hemolytic potency of peptides. *Sci Rep* 6:22843. <https://doi.org/10.1038/srep22843>
128. Win TS, Malik AA, Prachayasittikul V, JE SW, Nantasenamat C, Shoombuatong W (2017) HemoPred: a web server for predicting the hemolytic activity of peptides. *Future Med Chem* 9(3):275–291. <https://doi.org/10.4155/fmc-2016-0188>
129. Timmons PB, Hewage CM (2020) HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks. *Sci Rep* 10(1):10869. <https://doi.org/10.1038/s41598-020-67701-3>
130. Hasan MM, Schaduangrat N, Basith S, Lee G, Shoombuatong W, Manavalan B (2020) HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics* 36(11):3350–3356. <https://doi.org/10.1093/bioinformatics/btaa160>
131. Kedariseti P, Mizianty MJ, Kaas Q, Craik DJ, Kurgan L (2014) Prediction and characterization of cyclic proteins from sequences in three domains of life. *Biochim Biophys Acta* 1844:181–190. <https://doi.org/10.1016/j.bbapap.2013.05.002>

132. Agrawal P, Khater S, Gupta M, Sain N, Mohanty D (2017) RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. *Nucleic Acids Res* 45:W80–W88. <https://doi.org/10.1093/nar/gkx408>
133. Chen CH, Lu TK (2020) Development and challenges of antimicrobial peptides for therapeutic applications. *Antibiotics* 9:24. <https://doi.org/10.3390/antibiotics9010024>
134. Magana M, Pushpanathan M, Santos AL et al (2020) The value of antimicrobial peptides in the age of resistance. *Lancet Infect Dis* 20(9):E216–E230
135. Hancock REW, Patrzykat A (2002) Clinical Development of Cationic Antimicrobial Peptides: From Natural to Novel Antibiotics. *Current Drug Targets - Infectious Disorders* 2:79–83
136. Sheard DE, O'Brien-Simpson NM, Wade JD, Separovic F (2019) Combating bacterial resistance by combination of antibiotics with antimicrobial peptides. *Pure Appl Chem* 91(2):199–209
137. Fosgerau K, Hoffmann T (2015) Peptide therapeutics: current status and future directions. *Drug Discovery Today* 20:122–128
138. Werner HM, Cabaltea CC, Horne WS (2016) Peptide backbone composition and protease susceptibility: impact of modification type, position, and tandem substitution. *ChemBioChem* 17(8):712–718
139. Werner HM, Horne WS (2015) Folding and function in α/β -peptides: targets and therapeutic applications. *Curr Opin Chem Biol* 28:75–82



Molecular Mechanism of Action of Antimicrobial Agents Against Clinically Important Human Pathogens: A Proteomics Approach

16

Anthonyimuthu Selvaraj, Alaguvel Valliammai, and Shunmugiah Karutha Pandian

Abstract

Globally, prevalence of infectious diseases profoundly affects the human health and economy. The failure of conventional antimicrobial agents and emergence of antimicrobial resistance among the pathogens are the major reason for spreading of infectious diseases. Hence, the need for novel therapy is increased to control infectious diseases. Deciphering the mode of action of drug is an important and crucial process in novel drug discovery. The understanding of antimicrobial resistance mechanisms in pathogens is a vital task during drug development. In these aspects, proteomics provides an innovative platform for the understanding of alteration in protein pathways that are associated with antimicrobial resistance in pathogens. Further, proteomics study is also supporting to recognize how the drug kills pathogen and also to reveal drug targeting pathways of pathogens. To develop more efficient and novel therapies against pathogen infections, it is essential to study the pathogen's response to drugs and establish resistance mechanisms in pathogens. Proteomics is most suitable tool to unveil molecular mechanism of antimicrobial agents. In this chapter, we aimed to reveal the significance of proteomics-based approaches in the identification of antimicrobial drug targets, to decipher the mechanism behind the drug resistance, and to unveil the mode of action of antimicrobial agents.

Keywords

Infectious disease · Antimicrobial agents · Antimicrobial resistance · Proteomics · Mode of action

A. Selvaraj · A. Valliammai · S. K. Pandian (✉)
Department of Biotechnology, Alagappa University, Karaikudi, Tamil Nadu, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_16

287

16.1 Introduction

Normally, human microbiome is harmless and also supporting the several biological processes of host. Likewise, several pathogens are also having the ability to survive in humans without causing any harmful infections till the host is with active immune system. However, immune system of host is compromised at certain circumstances, and infectious agents take advantage of this condition to cause diseases. Infectious diseases are majorly caused by microorganisms such as bacteria, fungi, and viruses [1, 2]. Among the infectious disease-causing pathogens, bacteria and fungi play an important role in causing infectious diseases to human [3, 4]. Bacteria possess the ability to cause mild to severe infections including soft tissue infections, urinary tract infections, bacteremia, tuberculosis, bacterial meningitis, wound infections, pneumonia, etc. [5].

Antibiotics are predominantly used to control the bacterial infections and it works by completely killing bacteria or by affecting bacterial growth and development. Since the origin of antibiotics, they have significantly reduced the impact of infectious diseases caused by bacteria. At the same time, inappropriate usage of antibiotics led to the development of antibiotics resistance among the bacterial species, and it reduces the efficacy of antibiotics against infectious diseases [6]. Same phenomenon was observed in treatment of other pathogens associated with infectious diseases such as fungi and protozoans [7]. Antimicrobial resistance is causing serious public health problem by increasing the hospitalization, treatment costs, morbidity, and mortality. According to recent data of the Centers for Disease Control and Prevention (CDC), yearly more than 2.8 million cases are observed with antibiotic-resistant infections, and 35,000 people are dying annually in the USA due to infectious disease caused by multidrug-resistant pathogens. Furthermore, the cost spent for the treatment of infectious disease immensely affected the economic productivity up to \$1.5 billion per year in the USA. Overall, antimicrobial resistance impacted public health and economics worldwide [8–10]. Antimicrobial resistance is acquired by the pathogens through various mechanisms such as drug target modification, efflux pump, drug inactivation, biofilm formation, virulence factor secretion, etc. [11]. Therefore, emergence of antimicrobial resistance in pathogens is a global threat to human health and development. Hence, it requires more attention to develop novel antimicrobial therapy to treat infectious diseases caused by various pathogens.

In order to control antimicrobial resistance, numerous factors need to be considered during the development of novel antimicrobial agents including drug target, mechanism of action, drug efficacy, and safety. Among these factors, identification of drug targets and mechanism of action is a crucial aspect to avoid the development of antimicrobial resistance in pathogens. An important process in good drug designing is to identify the targets responsible for pathogen growth, development, and pathogenicity and designing drugs for vital druggable targets. The prediction of drug targets and mechanisms of action of drugs saves cost and time in the development of novel antimicrobial agents [11–14]. Elucidation of molecular mechanisms of drug has gained more importance for the development of new antimicrobial agents since it

reveals drug targeting pathways and thereby it supports the discovery of novel drug with multi-target potential against pathogens. Several approaches have been used to study drug targets and mechanism of action such as biochemical methods, computational methods, and omics-based methods including genomics-based method, deep sequencing method, and transcriptomics and proteomics approaches [15–17]. Among all these approaches, proteomics is the very effective method to identify drug target and molecular mechanism of action on infectious disease-causing pathogens. Since, proteins are real functional biomolecules in the living system, and they are involved in pathogenesis and resistance mechanism of pathogens [18–20]. In this regard, the current chapter briefs about importance of proteomics-based approaches to identify the molecular mechanism of action of antimicrobial agents against infectious disease-causing human pathogens.

16.2 Omics-Based Approaches to Analysis Mode of Action of Antimicrobial Agents

The omics-based approach is a true systematic approach to understand pathogen metabolism, drug resistance mechanism, and pathogenicity. Recently, omics-based approach is completely associated with drug discovery pipeline due to their role on drug target identification, drug target validation, druggable pathways, unraveling drug mode of actions, etc. [21, 22]. Genomics, transcriptomics, proteomics, and metabolomics are various omics strategies used in the novel drug discovery for microbial infections (Fig. 16.1) [23]. In genomics-based approach, whole genome sequencing helps antimicrobial drug discovery by exploring the virulence gene expression dynamism in pathogens grown in the absence and presence of novel antimicrobial agents of interest. And, it also reveals genes responsible for disruption of drug action, drug target modification, and efflux pump activations [24, 25]. Microarrays, RNA sequencing, and gene expression analyses belong to transcriptomics-based method to study the differential expression in organisms at various conditions. Transcriptomics analysis became an important method to understand mechanisms of pathogenicity and gene function, recognize new drug targets, and discover the drug action [26, 27]. Proteomics approach has attained more response in drug discovery due to limitations in the sensitivity and complexity of genomics and transcriptomics approaches. Proteomics techniques largely unveiled the proteins involved in pathogenesis, drug-resistant proteins, druggable targets, and pathways [28–30]. Proteomics-based approach plays unique and vital roles in the new drug development process because proteins are the real key players in living organisms.

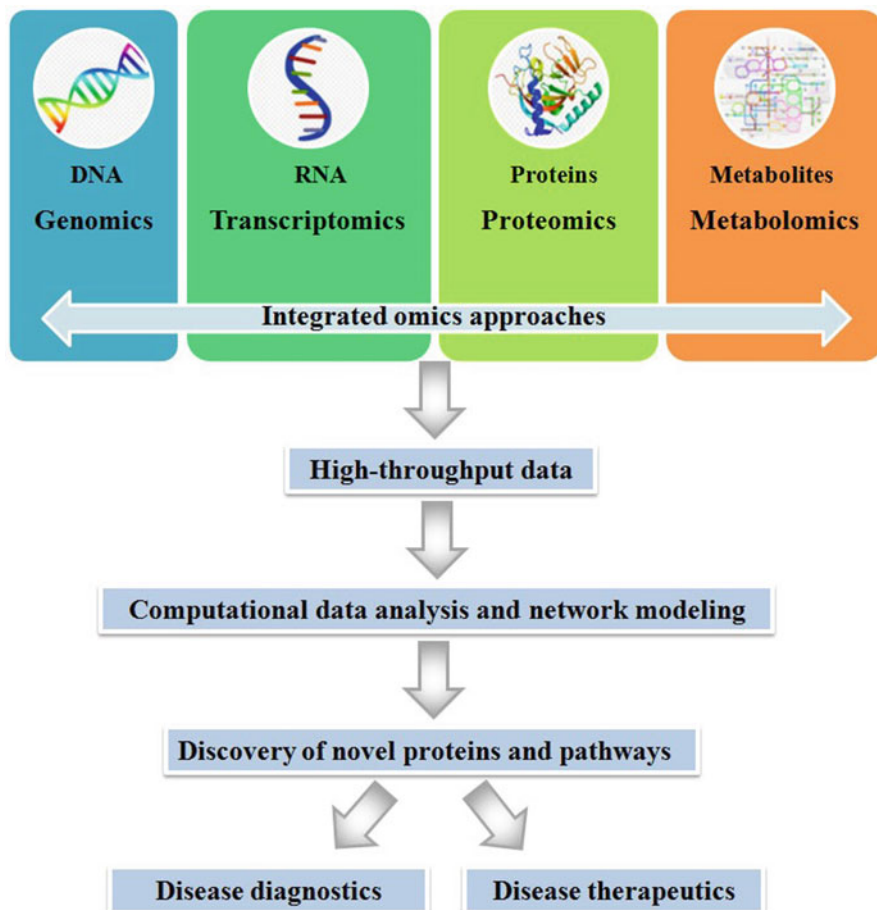


Fig. 16.1 Omics-based approaches in antimicrobial agent development

16.3 Proteomics and Its Significance

In 1994, Marc Wilkins coined the term proteome. Proteome is specifically studying the structures and functions of whole proteins of particular organisms. Proteins are important biomolecules of living organisms and are mediating all the metabolic pathways and biological process of organisms. Proteins are actual influencer in biological function and are not only dependent on DNA and mRNA expression levels but also with the posttranslational modification of host organisms [31–33]. Therefore, proteomics has been considered as the most suitable way to characterize biological systems when compared to genomics and transcriptomics. Proteomics is the technology used for the characterization, quantification, and identification of whole proteome of cell, tissue, or an organism. Proteomics-based approaches are

used in several aspects such as recognition of biomarkers, pathogenicity mechanisms, identification of differential expression of proteins, disease diagnosis, and elucidation of the role of proteins in various pathways of organisms [34, 35].

Proteomics technologies enable the identification of biomarkers for drug efficacy and toxicity, thereby helping the drug development process. Proteomics-based technologies are ideal choice for the identification of pathways targeted by novel drugs and revealed the function of proteins under disease conditions. It is useful to study the host-pathogen interactions, and thereby it supports to diagnosis of infectious diseases caused by pathogens. Proteomics experiments could be used for various purposes in clinical and health studies including monitoring the food proteins and biomarker discovery in various diseases such as tumor, AIDS, cardiovascular, and renal diseases [36–38]. Overall, the development and application of proteomics have been increased greatly in several promising new directions.

16.4 Proteomics in Drug Discovery

Currently, studying proteins of organism is an important process in drug discovery, and researchers have also focused on proteomics-based study to develop the novel drugs. Generally, application of proteomics includes identification and validation of drug target, identification of biomarkers to diagnose disease, assessment of toxicity of drugs, and mechanisms of action. Proteomics analyses reveal differential expression of proteins in response to infection-causing pathogens and thereby help to identify target proteins responsible for infections. These proteins could be a potential therapeutic target to design novel drugs [39–42].

Proteomics approaches also help to study the protein-protein interactions through which it supports to assess the impact of drugs on pathogens. Protein expression levels in pathogen are modified based on drug treatment and provide vital indications about drug effectiveness and targets. Target identification is the primary process in drug discovery, and also validation of identified targets is essential process in drug discovery pipeline [39, 40, 43–45]. Crucial process in drug development pipeline is to understand the virulence-associated pathways of microorganisms based on that drug need to be designed. Proteomics technologies are very useful to identify the pathways involved in pathogenesis of microorganisms. Proteomics-based technologies are suitable method to assess the drug resistance mechanisms in pathogens by comparing proteome of sensitive strains with resistance strains. Proteomics methods are also useful to identify posttranslational protein modifications including phosphorylation, glycosylation, acetylation, proteolysis, and amino acid polymorphisms in organisms [46–48]. On the whole, proteomics-based technologies are playing crucial role in terms of drug-target interaction, drug efficacy, drug toxicity, exposing the drug mechanism of action, drug resistance, etc.

16.5 Proteomics-Based Techniques for Studying Drug Development

Proteomics-based approaches have emerged as a powerful tool to study mechanism of action of particular drug. Figure 16.2 illustrates the various methods used in proteomics-based approaches. The top-down proteomics assess the proteins and posttranslational modifications in the intact state. The bottom-up strategy is sensitive and powerful approach to examine multiple proteins in a single sample and majorly used in clinical diagnostics. In bottom-up proteomics, proteins are enzymatically digested into small peptides, whereas in middle-down approach, proteins are digested into large peptides. The peptide fingerprint of the proteins is identified using liquid chromatographic pre-fractionation followed by mass spectrometry. But in top-down proteomics approach, the complex protein mixtures are ionized, fragmented, and analyzed in the intact form to identify the targeted proteins of the respective organisms [49].

Various proteomics analytical methods have been used to assess the drug mechanisms of action such as gel-based and gel-free method. In gel-based, proteins are extracted from pathogens grown without and with drugs. Then, proteins are separated based on their isoelectric point and molecular weight using two-dimensional gel electrophoresis. Then, the differentially expressed protein spots on the gels are selected and identified with the help of mass spectrometry

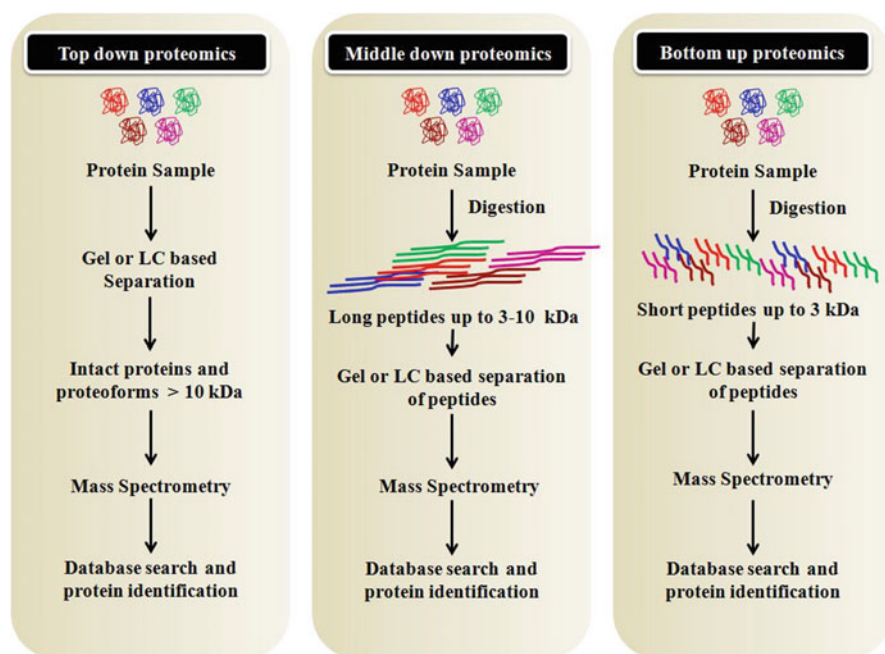


Fig. 16.2 Various strategies in proteomics-based approaches

analysis. On the other hand, in gel free methods, protein samples are subjected to mass spectrometric analysis without gel based separation. Various quantitative gel free proteomics technologies such as SILAC, iTRAQ and ICAT are widely used by researchers to identify differentially regulated proteins in crude samples [50–52]. Finally, differentially expressed proteins in drug-treated and untreated control proteins of organism could be found, and then drug-targeted pathways could be identified with computational analysis.

16.6 Infectious Disease-Causing Clinically Important Pathogens

Pathogen is defined as microorganism which possesses the ability to infect the host organisms. Mostly, human is an ideal choice for the pathogens because the nutritional availability and optimum temperature of the human body support the pathogen survival and multiplication. Numerous factors are responsible for spreading of infectious disease including global warming, urbanization, lifestyle, and inappropriate usage of antimicrobial agents [53–55]. Even though there is a remarkable advancement in the prevention, diagnosis, and treatment, infectious diseases still remain the leading cause of morbidity and mortality around the world. The resistance of pathogens to various antimicrobial agents has emerged as a major threat to the public health due to reduced efficacy of antimicrobial agents in the treatment of infectious diseases. Almost all the pathogens (bacteria, fungi, and virus) have high levels of multidrug resistance to conventional drugs. The development of drug resistance is a natural phenomenon due to inappropriate usage of antimicrobial drugs. Improper infection prevention and treatment led to emergence of drug resistance in pathogens [56–58]. Hence, it is more important to research more on the discovery of novel antimicrobial agent against infectious disease-causing pathogens.

Several studies have reported the various bacteria with high rates of infection such as *Acinetobacter* species, *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Salmonella*, *Shigella* species, *Neisseria*, and *Mycobacterium tuberculosis*. A high number of fungal infections are caused by *Trichosporon beigelii*, *Cryptococcus neoformans*, *Pseudallescheria boydii*, *Aspergillus* species, *Scopulariopsis* species, and *Candida* species. Prolonged viral infections have been observed in HIV virus, cytomegalovirus, herpes simplex virus, influenza A virus, varicella-zoster virus, hepatitis C, and SARS. *Plasmodium* species, *Toxoplasma gondii*, *Leishmania* species, *Trichomonas vaginalis*, and *Entamoeba* species are major disease-causing parasites [57, 59–61]. Table 16.1 presents the list of major infectious disease with causative agents.

Table 16.1 Major infectious diseases in humans

Type	Causative agent	Disease
Viral diseases	Influenza virus	Influenza
	Human immunodeficiency virus (HIV)	Acquired immunodeficiency syndrome (AIDS)
	Human papillomavirus (HPV)	Genital warts
	Hepatitis virus	Hepatitis
	Corona virus	Coronavirus disease (COVID)
Bacterial diseases	<i>Mycobacterium tuberculosis</i>	Tuberculosis
	<i>Salmonella typhi</i>	Typhoid
	<i>Helicobacter pylori</i>	Gastritis and ulcers
	<i>Neisseria gonorrhoeae</i>	Gonorrhoea
	<i>Neisseria meningitidis</i>	Meningitis
	<i>Staphylococcus aureus</i>	Toxic shock syndrome and soft tissue skin infections
	<i>Streptococcus pneumoniae</i>	Pneumonia
	<i>Streptococcus pyogenes</i>	Scarlet fever and strep throat
	<i>Clostridium tetani</i>	Tetanus
	<i>Corynebacterium diphtheriae</i>	Diphtheria
<i>Vibrio cholerae</i>	Cholera	
Fungal diseases	<i>Candida</i> spp.	Candidiasis
	<i>Cryptococcus neoformans</i>	Cryptococcal meningitis
	<i>Trichophyton</i> , <i>Microsporium</i> , and <i>Epidermophyton</i> spp.	Ringworm
	<i>Aspergillus</i> spp.	Aspergillosis
	<i>Blastomyces dermatitidis</i>	Pulmonary blastomycosis
Protozoan diseases	<i>Plasmodium</i> spp.	Malaria
	<i>Leishmania</i> spp.	Leishmaniasis
	<i>Trypanosoma</i> spp.	African trypanosomiasis/African sleeping sickness
	<i>Trypanosoma cruzi</i>	Chagas disease/American trypanosomiasis
	<i>Toxoplasma gondii</i>	Toxoplasmosis

16.7 Mode of Action of Antimicrobial Agents Elucidated Through Proteomics-Based Approaches

For many decades, antimicrobial agents are used to eradicate the infectious diseases. However, these antimicrobial agents became ineffective against pathogens due to various mechanisms such as efflux pump, biofilm, virulence factors, alteration of drug targets, and drugs degradation [62–64]. Hence, to develop more proficient antimicrobial agents against infectious diseases, studying the mechanism of action of drugs is a primary process. Proteomics is an appropriate and powerful tool to study molecular response of pathogens to antimicrobial compounds. Analyzing the

pathogen response profiling in the presence of drug could reveal the mechanisms behind resistance and tolerance of pathogens to antimicrobial agents [29, 65]. In addition, another importance of proteomics analyses in drug discovery process is to verify the activity of particular drug by checking their efficacy with drug targets.

Recently, several studies have reported the mechanism of action of antimicrobial agents on pathogens and also validated the drug targets using proteomics-based approaches [66–68]. In previous studies, proteome of pathogens grown in the presence and absence of antimicrobial agent of interest is evaluated by assessing the changes in protein expression level using proteomics techniques. The previous study on *P. aeruginosa* identified that curcumin altered the expression of proteins involved in iron acquisition, pyoverdine and pyocyanin production in *P. aeruginosa* to inhibit biofilm and virulence factors [69]. Another proteomics-based study reported that the antibiofilm agent citral inhibited the biofilm and virulence of *S. aureus* by affecting the expression of IsaA, CodY, and SaeS [70]. Further, proteomics-based study unraveled the variation in ergosterol, sphingolipid, and oxidative stress systems in *Candida albicans* by antifungal agent myristic acid [71]. Various studies successfully explored the mechanism of action of identified drugs on pathogens using proteomics-based tools (Table 16.2).

16.8 Conclusion

The rise of antimicrobial resistance in microbial pathogens has become major problem in an international scale. Therefore, there is an everlasting need for the development of new antimicrobial agents and also need to increase their therapeutic potential through by understanding the drug molecular mechanisms of action. The scientific community started to reveal the drug actions on pathogens and also decipher the resistance and tolerance development to antimicrobial agents. The proteome approaches not only support the drug discovery but also improve novel strategies to infectious disease-causing pathogens. This chapter summarized the importance of proteomics-based approaches for understanding mode of action of antimicrobial agents, role of proteomics in drug discovery, and support of proteomics to develop novel treatment strategies to control infectious disease.

Acknowledgment The authors sincerely acknowledge DST-FIST [Grant No. SR/FST/LSI-639/2015(C)], UGC-SAP [Grant No. F.5-1/2018/DRS-II(SAP-II)] and DST-PURSE [Grant No. SR/PURSE Phase 2/38 (G)] for providing instrumentation facilities. The authors also thank RUSA 2.0 [F.24-51/2014-U, Policy (TN Multi-Gen), Dept. of Edn, GoI], and SKP is thankful to UGC for Mid-Career Award [F.19-225/2018(BSR)].

Competing Interest

All the authors declare no conflict of interest.

Table 16.2 Mechanism of action of bioactives revealed by proteomics approaches

S. No.	Drug/compound	Target pathogen	Molecular mechanism	Reference
1.	Curcumin	<i>Pseudomonas aeruginosa</i>	Modulation of iron homeostasis and oxidative stress response	[69]
2.	Oleic acid	<i>Candida albicans</i>	Down regulation of ergosterol biosynthesis, lipase production, iron homeostasis	[66]
3.	Myristic acid	<i>Candida albicans</i>	Alteration of ergosterol, sphingolipid and oxidative stress	[71]
4.	Citral	<i>Acinetobacter baumannii</i>	Interruption of antibiotic resistance, antioxidant defense, and biofilm-associated two-component systems	[67]
5.	Vanillic acid	<i>Serratia marcescens</i>	S-layer, flagellin, and fatty acid biosynthesis	[68]
6.	Citral	<i>Staphylococcus aureus</i>	Modulation of pleotropic transcriptional repressor, cell wall homeostasis, exotoxin secretion	[70]
7.	3-p-Trans-coumaroyl-2-hydroxyquinic acid	<i>Staphylococcus aureus</i>	Disruption of cell membrane and peptidoglycan synthesis	[72]
8.	Bismuth drugs	<i>Helicobacter pylori</i>	Urease accessory protein ureg	[73]
9.	Chlorhexidine	<i>Acinetobacter baumannii</i>	Disruption of cell membrane	[74]
10.	Silver nanoparticles	<i>Pseudomonas aeruginosa</i>	Stimulation of oxidative stress response, an destroying iron homeostasis	[75]
11.	3-Hydroxyphenylacetic acid	<i>Pseudomonas aeruginosa</i>	Modulation of DNA replication and repair, RNA modifications	[76]
12.	Plantaricin gz1-27	<i>Staphylococcus aureus</i>	Modulation of biofilm formation, DNA replication and repair, and heat-shock	[77]
13.	Chitosan	<i>Escherichia coli</i>	Altering the stability of outer membrane	[78]
14.	Silver	<i>Escherichia coli</i>	Damage of multiple enzymes in glycolysis and tricarboxylic acid (tca) cycle	[79]
15.	Alpha-mangostin	<i>Staphylococcus epidermidis</i>	Alteration in cytoplasmic membrane integrity, cell division, teichoic acid biosynthesis	[80]

(continued)

Table 16.2 (continued)

S. No.	Drug/compound	Target pathogen	Molecular mechanism	Reference
16.	Daptomycin	<i>Staphylococcus aureus</i>	Disruption of cell membrane	[81]
17.	Rhodomyrtone	<i>Staphylococcus aureus</i>	Disruption of cell wall biosynthesis and cell division	[82]
18.	Gold nanoparticles	<i>Escherichia coli</i>	Modulation of energy metabolism and transcription	[83]
19.	Zinc oxide nanoparticle	<i>Acinetobacter baumannii</i>	Production of reactive oxygen species and membrane leakage	[84]
20.	B-hairpin macrocyclic peptide jb-95	<i>Escherichia coli</i>	Targeting outer membrane proteins	[85]

References

- Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, Bhatt S, Brownstein JS, Collier N, Myers MF, George DB, Gething PW (2013) Global mapping of infectious disease. *Phil Trans R Soc B Biol Sci* 368(1614):20120250
- Vynnycky E, White R (2010) An introduction to infectious disease modelling. OUP, Oxford
- Peleg AY, Hogan DA, Mylonakis E (2010) Medically important bacterial–fungal interactions. *Nat Rev Microbiol* 8(5):340–349
- Wargo MJ, Hogan DA (2006) Fungal–bacterial interactions: a mixed bag of mingling microbes. *Curr Opin Microbiol* 9(4):359–364
- Doron S, Gorbach SL (2008) Bacterial infections: overview. *International Encyclopedia of Public Health*, p. 273
- Ventola CL (2015) The antibiotic resistance crisis: part 1: causes and threats. *Pharm Therap* 40(4):277
- Pai V, Ganavalli A, Kikkeri NN (2018) Antifungal resistance in dermatology. *Indian J Dermatol* 63(5):361
- Campos JCDM, Antunes LC, Ferreira RB (2020) Global priority pathogens: virulence, antimicrobial resistance and prospective treatment options. *Future Microbiol* 15:649–677
- Martins AF, Rabinowitz P (2020) The impact of antimicrobial resistance in the environment on public health. *Future Microbiol* 15:699–702
- Centres for Disease Control and Prevention (US) (2013) Antibiotic resistance threats in the United States, 2013. Centres for Disease Control and Prevention, US Department of Health and Human Services
- Reygaert WC (2018) An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS Microbiol* 4(3):482
- Belete TM (2019) Novel targets to develop new antibacterial agents and novel alternatives to antibacterial agents. *Human Microb J* 11:100052
- Brown JR (2010) The Design and synthesis of novel antimicrobial agents for use in the battle against bacterial resistance. Theses and Dissertations (ETD). Paper 31. <https://doi.org/10.21007/etd.cghs.2010.0035>
- Lohner K (2001) Development of novel antimicrobial agents: emerging strategies. Horizon Scientific Press, Wymondham

15. Misra BB, Langefeld C, Olivier M, Cox LA (2019) Integrated omics: tools, advances and future approaches. *J Mol Endocrinol* 62(1):R21–R45
16. Mohana NC, Rao HY, Rakshith D, Mithun PR, Nuthan BR, Satish S (2018) Omics based approach for biodiscovery of microbial natural products in antibiotic resistance era. *J Genet Eng Biotechnol* 16(1):1–8
17. Wolfender JL, Litaudon M, Touboul D, Queiroz EF (2019) Innovative omics-based approaches for prioritisation and targeted isolation of natural products – new strategies for drug discovery. *Nat Prod Rep* 36(6):855–868
18. Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH (2017) Proteomics: technologies and their applications. *J Chromatogr Sci* 55(2):182–196
19. Cho WC (2007) Proteomics technologies and challenges. *Genomics Proteomics Bioinformatics* 5(2):77–85
20. Yoithappabhunath TR, Nirmal RM, Santhadevy A, Anusushanth A, Charanya D (2015) Role of proteomics in physiologic and pathologic conditions of dentistry: overview. *J Pharm Bioallied Sci* 7(Suppl 2):S344
21. Cowell AN, Winzeler EA (2019) Advances in omics-based methods to identify novel targets for malaria and other parasitic protozoan infections. *Genome Med* 11(1):63
22. Pelkonen O, Pasanen M, Lindon JC, Chan K, Zhao L, Deal G, Xu Q, Fan TP (2012) Omics and its potential impact on R&D and regulation of complex herbal products. *J Ethnopharmacol* 140(3):587–593
23. Hasin Y, Seldin M, Lusic A (2017) Multi-omics approaches to disease. *Genome Biol* 18(1):1–15
24. Hall J, Dennler P, Haller S, Pratsinis A, Sauberli K, Towbin H, Walthe K, Woytschak J (2010) Genomics drugs in clinical trials. *Nat Rev Drug Discov* 9(12):988
25. Mills SD (2003) The role of genomics in antimicrobial discovery. *J Antimicrob Chemother* 51(4):749–752
26. Domínguez Á, Muñoz E, López MC, Cordero M, Martínez JP, Viñas M (2017) Transcriptomics as a tool to discover new antibacterial targets. *Biotechnol Lett* 39(6):819–828
27. O'Rourke A, Beyhan S, Choi Y, Morales P, Chan AP, Espinoza JL, Dupont CL, Meyer KJ, Sporing A, Lewis K, Nierman WC (2020) Mechanism-of-action classification of antibiotics by global transcriptome profiling. *Antimicrob Agents Chemother* 64(3):e01207–19
28. Goff A, Cantillon D, Muraro Wildner L, Waddell SJ (2020) Multi-omics technologies applied to tuberculosis drug discovery. *Appl Sci* 10(13):4629
29. Khodadadi E, Zeinalzadeh E, Taghizadeh S, Mehramouz B, Kamounah FS, Khodadadi E, Ganbarov K, Yousefi B, Bastami M, Kafil HS (2020) Proteomic applications in antimicrobial resistance and clinical microbiology studies. *Infect Drug Resist* 13:1785
30. Pérez-Llarena FJ, Bou G (2016) Proteomics as a tool for studying bacterial virulence and antimicrobial resistance. *Front Microbiol* 7:410
31. Ahmad Y, Lamond AI (2014) A perspective on proteomics in cell biology. *Trends Cell Biol* 24(4):257–264
32. Graves PR, Haystead TA (2002) Molecular biologist's guide to proteomics. *Microbiol Mol Biol Rev* 66(1):39–63
33. Wen B, Zeng WF, Liao Y, Shi Z, Savage SR, Jiang W, Zhang B (2020) Deep learning in proteomics. *Proteomics* 20(21–22):1900335
34. Burbaum J, Tobal GM (2002) Proteomics in drug discovery. *Curr Opin Chem Biol* 6(4):427–433
35. Yosida M, Loo JA, Lepley RA (2001) Proteomics as a tool in the pharmaceutical drug design process. *Curr Pharm Des* 7(4):291–310
36. Boteanu RM, Suica VI, Ivan L, Safciuc F, Uyy E, Dragan E, Croitoru SM, Grumezescu V, Chiritoiu M, Sima LE, Vlagioiu C (2020) Proteomics of regenerated tissue in response to a titanium implant with a bioactive surface in a rat tibial defect model. *Sci Rep* 10(1):1–13

37. Corbo C, Parodi A, Evangelopoulos M, Engler D, K Matsunami R, C Engler A, Molinaro R, Scaria S, Salvatore F, Tasciotti E (2015) Proteomic profiling of a biomimetic drug delivery platform. *Curr Drug Targets* 16(13):1540–1547
38. Hedl TJ, San Gil R, Cheng F, Rayner SL, Davidson JM, De Luca A, Villalva MD, Ecroyd H, Walker AK, Lee A (2019) Proteomics approaches for biomarker and drug target discovery in ALS and FTD. *Front Neurosci* 13:548
39. Ryan TE, Patterson SD (2002a) Proteomics in drug target discovery: high-throughput meets high-efficiency. *Drug Discov World* 3:43–52
40. Ryan TE, Patterson SD (2002b) Proteomics: drug target discovery on an industrial scale. *Trends Biotechnol* 20(12):s45–s51
41. Veenstra TD (2006) Proteomic approaches in drug discovery. *Drug Discov Today Technol* 3(4):433–440
42. Walgren JL, Thompson DC (2004) Application of proteomic technologies in the drug development process. *Toxicol Lett* 149(1–3):377–385
43. Jain KK (2002) Proteomics-based anticancer drug discovery and development. *Technol Cancer Res Treat* 1(4):231–236
44. Jain KK (2004) Applications of proteomics technologies for drug discovery. In: *Proteomics: biomedical and pharmaceutical applications*. Springer, Dordrecht, pp 201–227
45. Jain KK (2005) Proteomics-based anticancer drug discovery. In: *The oncogenomics handbook*. Humana Press, Totowa, pp 123–134
46. Li Y, Wu T (2018) Proteomic approaches for novel systemic lupus erythematosus (SLE) drug discovery. *Expert Opin Drug Discovery* 13(8):765–777
47. Roti G, Stegmaier K (2012) Genetic and proteomic approaches to identify cancer drug targets. *Br J Cancer* 106(2):254–261
48. Verhelst SH, Bogoyo M (2005) Chemical proteomics applied to target identification and drug discovery. *BioTechniques* 38(2):175–177
49. Wehr T (2006) Top-down versus bottom-up approaches in proteomics. *Lc Gc North America* 24:9
50. Bantscheff M, Hopf C, Kruse U, Drewes G (2008) Proteomics-based strategies in kinase drug discovery. In: *Sparkling signals*. Springer, Berlin, pp 1–28
51. Schirle M, Bantscheff M, Kuster B (2012) Mass spectrometry-based proteomics in preclinical drug discovery. *Chem Biol* 19(1):72–84
52. Wright PC, Noirel J, Ow SY, Fazeli A (2012) A review of current proteomics technologies with a survey on their widespread use in reproductive biology investigations. *Theriogenology* 77(4):738–765
53. Brachman PS (2003) Infectious diseases—past, present, and future. *Int J Epidemiol* 32(5):684–686
54. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P (2008) Global trends in emerging infectious diseases. *Nature* 451(7181):990–993
55. Ndow G, Ambe JR, Tomori O (2019) Emerging infectious diseases: a historical and scientific review. In: *Socio-cultural dimensions of emerging infectious diseases in Africa*. Springer, Cham, pp 31–40
56. Morens DM, Folkers GK, Fauci AS (2010) The challenge of emerging and re-emerging infectious diseases. *Nature* 463(7277):122
57. Nii-Trebi NI (2017) Emerging and neglected infectious diseases: insights, advances, and challenges. *Biomed Res Int* 2017:5245021
58. Raoult D, Roux V (1997) Rickettsioses as paradigms of new or emerging infectious diseases. *Clin Microbiol Rev* 10(4):694–719
59. Boggild AK, Libman M, Greenaway C, McCarthy AE, and Committee to Advise on Tropical Medicine (2016) Emerging infectious diseases: CATMAT statement on disseminated strongyloidiasis: prevention, assessment and management guidelines. *Can Commun Dis Rep* 42(1):12
60. Hughes JM (2001) Emerging infectious diseases: a CDC perspective. *Emerg Infect Dis* 7(3 Suppl):494

61. Mehand MS, Al-Shorbaji F, Millett P, Murgue B (2018) The WHO R&D Blueprint: 2018 review of emerging infectious diseases requiring urgent research and development efforts. *Antivir Res* 159:63–67
62. Christaki E, Marcou M, Tofarides A (2020) Antimicrobial resistance in bacteria: mechanisms, evolution, and persistence. *J Mol Evol* 88(1):26–40
63. Morrison L, Zembower TR (2020) Antimicrobial resistance. *Gastrointest Endosc Clin N Am* 30(4):619–635
64. Schrader SM, Vaubourgeix J, Nathan C (2020) Biology of antimicrobial resistance and approaches to combat it. *Sci Transl Med* 12(549):eaz6992
65. Tsakou F, Jersie-Christensen R, Jenssen H, Mojsoska B (2020) The role of proteomics in bacterial response to antibiotics. *Pharmaceuticals* 13(9):214
66. Muthamil S, Prasath KG, Priya A, Precilla P, Pandian SK (2020) Global proteomic analysis deciphers the mechanism of action of plant derived oleic acid against *Candida albicans* virulence and biofilm formation. *Sci Rep* 10(1):1–17
67. Selvaraj A, Valliammai A, Muthuramalingam P, Sethupathy S, Subramenium GA, Ramesh M, Pandian SK (2020) Proteomic and systematic functional profiling unveils citral targeting antibiotic resistance, antioxidant defense, and biofilm-associated two-component systems of *Acinetobacter baumannii* to cucumber biofilm and virulence traits. *mSystems* 5(6):e00986–20
68. Sethupathy S, Ananthi S, Selvaraj A, Shanmuganathan B, Vigneshwari L, Balamurugan K, Mahalingam S, Pandian SK (2017) Vanillic acid from *Actinidia deliciosa* impedes virulence in *Serratia marcescens* by affecting S-layer, flagellin and fatty acid biosynthesis proteins. *Sci Rep* 7(1):1–17
69. Sethupathy S, Prasath KG, Ananthi S, Mahalingam S, Balan SY, Pandian SK (2016) Proteomic analysis reveals modulation of iron homeostasis and oxidative stress response in *Pseudomonas aeruginosa* PAO1 by curcumin inhibiting quorum sensing regulated virulence factors and biofilm production. *J Proteome* 145:112–126
70. Valliammai A, Sethupathy S, Ananthi S, Priya A, Selvaraj A, Nivetha V, Aravindraja C, Mahalingam S, Pandian SK (2020) Proteomic profiling unveils citral modulating expression of IsaA, CodY and SaeS to inhibit biofilm and virulence in methicillin-resistant *Staphylococcus aureus*. *Int J Biol Macromol* 158:208–221
71. Prasath KG, Sethupathy S, Pandian SK (2019) Proteomic analysis uncovers the modulation of ergosterol, sphingolipid and oxidative stress pathway by myristic acid impeding biofilm and virulence in *Candida albicans*. *J Proteome* 208:103503
72. Liu X, Yue Y, Wu Y, Zhong K, Bu Q, Gao H (2020) Discovering the antibacterial mode of action of 3-p-trans-coumaroyl-2-hydroxyquinic acid, a natural phenolic compound, against *Staphylococcus aureus* through an integrated transcriptomic and proteomic approach. *J Food Saf* 41:e12861
73. Li H, Wang R, Sun H (2018) Systems approaches for unveiling the mechanism of action of bismuth drugs: new medicinal applications beyond *Helicobacter pylori* infection. *Acc Chem Res* 52(1):216–227
74. Biswas D, Tiwari M, Tiwari V (2019) Molecular mechanism of antimicrobial activity of chlorhexidine against carbapenem-resistant *Acinetobacter baumannii*. *PLoS One* 14(10):e0224107
75. Zhang Y, Pan X, Liao S, Jiang C, Wang L, Tang Y, Wu G, Dai G, Chen L (2020) Quantitative proteomics reveals the mechanism of silver nanoparticles against multidrug-resistant *Pseudomonas aeruginosa* biofilms. *J Proteome Res* 19(8):3109–3122
76. Ozdemir OO, Soyer F (2020) *Pseudomonas aeruginosa* presents multiple vital changes in its proteome in the presence of 3-hydroxyphenylacetic acid, a promising antimicrobial agent. *ACS Omega* 5(32):19938–19951
77. Du H, Zhou L, Lu Z, Bie X, Zhao H, Niu YD, Lu F (2020) Transcriptomic and proteomic profiling response of methicillin-resistant *Staphylococcus aureus* (MRSA) to a novel bacteriocin, plantaricin GZ1-27 and its inhibition of biofilm formation. *Appl Microbiol Biotechnol* 104(18):7957–7970

78. Gomes LP, Anjo SI, Manadas B, Coelho AV, Paschoalin VM (2020) Proteomic analyses reveal new insights on the antimicrobial mechanisms of chitosan biopolymers and their nanosized particles against *Escherichia coli*. *Int J Mol Sci* 21(1):225
79. Wang H, Yan A, Liu Z, Yang X, Xu Z, Wang Y, Wang R, Koohi-Moghadam M, Hu L, Xia W, Tang H (2019) Deciphering molecular mechanism of silver by integrated omic approaches enables enhancing its antimicrobial efficacy in *E. coli*. *PLoS Biol* 17(6):e3000292
80. Sivaranjani M, Leskinen K, Aravindraja C, Saavalainen P, Pandian SK, Skurnik M, Ravi AV (2019) Deciphering the antibacterial mode of action of alpha-mangostin on *Staphylococcus epidermidis* RP62A through an integrated transcriptomic and proteomic approach. *Front Microbiol* 10:150
81. Ma W, Zhang D, Li G, Liu J, He G, Zhang P, Yang L, Zhu H, Xu N, Liang S (2017) Antibacterial mechanism of daptomycin antibiotic against *Staphylococcus aureus* based on a quantitative bacterial proteome analysis. *J Proteome* 150:242–251
82. Sianglum W, Srimanote P, Wonglumsom W, Kittiniyom K, Voravuthikunchai SP (2011) Proteome analyses of cellular proteins in methicillin-resistant *Staphylococcus aureus* treated with rhodomyrtone, a novel antibiotic candidate. *PLoS One* 6(2):e16628
83. Cui Y, Zhao Y, Tian Y, Zhang W, Lü X, Jiang X (2012) The molecular mechanism of action of bactericidal gold nanoparticles on *Escherichia coli*. *Biomaterials* 33(7):2327–2333
84. Tiwari V, Mishra N, Gadani K, Solanki PS, Shah NA, Tiwari M (2018) Mechanism of antibacterial activity of zinc oxide nanoparticle against carbapenem-resistant *Acinetobacter baumannii*. *Front Microbiol* 9:1218
85. Urfer M, Bogdanovic J, Monte FL, Moehle K, Zerbe K, Omasits U, Ahrens CH, Pessi G, Eberl L, Robinson JA (2016) A peptidomimetic antibiotic targets outer membrane proteins and disrupts selectively the outer membrane in *Escherichia coli*. *J Biol Chem* 291(4):1921–1932



Exploration of the Mycobacterial Proteome in the Pathogenesis of TB: A Perspective 17

Mohd. Shariq, Sheeba Zarin, Nilisha Rastogi, Indu Kumari, Farha Naz, Tarina Sharma, Neha Sharma, and Nasreen Z. Ehtesham

Abstract

Tuberculosis caused by *Mycobacterium tuberculosis* (*M. tb*) with a global burden of 10 million cases in 2019 causes morbidity and mortality if not treated. The constant rise in drug-resistant TB has further aggravated the problems and proves to be the major roadblock in combatting TB. Complete understanding of the pathogen's physiology and its virulence attributes is essentially required and is important in designing new treatment strategies. *M. tb* multi-omics strategies are proving to be very useful in gaining insights into the disease. Despite the availability of genomic and transcriptomic data, the pathogenic potential, survival strategies, persistence, immunomodulation, drug resistance mechanisms and the host-pathogen interactions remain poorly understood. Proteomics approaches thus prove to be more informative for studying mycobacteria providing minute details about the true state of the cell under various conditions. Protein profiling of different strains of mycobacteria, clinically relevant, as well as drug-resistant isolates, has tremendously increased our knowledge in the understanding of disease mechanism. Proteomics and bioinformatics approaches have helped greatly in identification and characterization of target proteins which can be exploited for novel therapeutics. This book chapter provides an update on the different proteomics technologies and their application in unravelling TB physiology.

Keywords

M. tuberculosis · Drug resistant · Drug discovery · Proteomic analysis · X-ray crystallography

M. Shariq · S. Zarin · N. Rastogi · I. Kumari · F. Naz · T. Sharma · N. Sharma · N. Z. Ehtesham (✉)
Inflammation Biology and Cell Signalling Laboratory, ICMR National Institute of Pathology, New Delhi, India

17.1 Introduction

Mycobacterium tuberculosis (*M. tb*), the aetiological agent for tuberculosis (TB), is ranked as one of the topmost killers among all infectious agents [1]. The major challenges for the management of TB include the rise of multidrug-resistant (MDR) TB, extensively drug-resistant TB (XDR), HIV (human immunodeficiency virus) co-infection and poverty [2]. There is no effective vaccine, and the only vaccine available is the century-old Bacillus Calmette-Guerin (BCG). However, the variable protection of BCG in adults has posed a serious threat to TB elimination programme worldwide [3]. The current therapy comprises first-line drugs, which take 6 to 9 months for completion and have serious side effects [4]. Treatment of MDR/XDR requires much longer duration and includes TB drugs from the second line in addition to first-line drugs, pyrazinamide and a high dose of isoniazid [5]. Despite these treatment regimes, the rise in MDR and XDR cases has created new hurdles for pre-existing drug therapy [6]. Hence, there is an unmet medical need to develop an effective vaccine and improved drugs for TB management.

M. tb is transmitted through inhalation of airborne aerosol bearing the pathogen [7]. Upon entry of *M. tb* pathogen into the lungs, it infects alveolar macrophages and bypasses the host immune response for its survival and pathogenesis [8]. To further ensure its survival, mycobacteria also dampen anti-mycobacterial defence mechanism utilised by macrophages including autophagy, phagosome acidification and production of reactive oxygen and nitrogen species [9–11]. Also, infected alveolar macrophages then secrete chemokines that attract inflammatory cells including neutrophils, macrophages and natural killer cells, further promoting inflammation and formation of multinucleated giant cells called granulomas [12]. These granulomas thus provide a niche for the containment of bacteria and also serve as a reservoir for the spread of the infection.

M. tb-secreted proteins (secretome) play a critical role in subverting the immune response and intracellular growth [13, 14]. Early secreted antigen (ESAT-6), an essential virulence factor of *M. tb*, is known for regulation of host immune response by inhibiting pro-inflammatory responses, such as interferon (IFN) gamma production [15] and interleukin (IL)-12 production [16]. Furthermore, ESAT-6 stimulates IL-6 production in macrophages [17]. Besides, ESAT-6 also plays an important role in inducing macrophage polarisation and transition into epithelioid macrophages, the major constituent of TB granuloma [18, 19]. It was further demonstrated that *M. tb* secreted effector Rv1988 localises to the host nucleus and methylate host histone proteins and thus epigenetically modulates macrophage's anti-mycobacterial functions [20]. These studies cumulatively suggest the critical role of *M. tb* proteins in the regulation of host functions, and hence, for the understanding of *M. tb*-associated disease pathology, characterisation of *M. tb* proteome is warranted [21]. In addition to characterising the role of *M. tb* proteins in virulence, the *M. tb* proteome could also be explored for potential antigens that can be utilised as an effective vaccine candidate. In this regard, high-throughput proteome-wide screening of potential *M. tb* antigens could be helpful in the generation of novel vaccines [22].

Although the genomic makeup of *M. tb* has been extensively studied, the proteome analysis of *M. tb* lags due to the complex protocols for *M. tb* protein isolation and need for sophisticated instrumentation [23]. Also, the proteome of XDR *M. tb* revealed that more than 30% are hypothetical proteins that have not been characterised to date [24]. The lack of comprehensive exploration of *M. tb* proteome has further widened the gap in understanding the virulence and pathogenesis of *M. tb*. Therefore, unravelling the proteomic makeup of *M. tb* will further be helpful in better understanding of the physiology and virulence of *M. tb*, which may reveal novel drug targets [23]. Altogether these efforts may help achieve the WHO's End TB strategy. *M. tb* H37Rv complete genome of 4.4 Mb consists of around 3924 genes [25]. The identification and characterisation of all the genes are important, but attention should be given to the gene products responsible for virulence and pathogenic attributes. These virulence factors can then be quantitated using proteomics approach amounting for the difference in pathogenicity and drug resistance among lineages.

17.2 Investigation of Mycobacteria Using Proteomics Approach

Proteomics proves to be an important tool in the identification of novel protein targets which are part of pathogen survival strategies, defence responses by the host and subsequently, the host-pathogen interactions. The upregulation and downregulation of host immune-related proteins and virulence-related proteins of pathogens are indicative of their role in defence or pathogenesis. These upregulation and downregulation are useful in the identification of proteins which may prove to be important as drug targets or development of diagnostic tools representing various stages of the pathogenesis and the level of advancement of the infection. Starting with the identification of any such protein to establishing it as a drug target or a diagnostic marker and to monitor the kinetics of protein contents of different organs in response to infections, we need to proceed with an approach involving the following steps in a sequential pattern:

Identification of novel targets

1. Proteomic analysis
2. Bioinformatic analysis
3. Biochemical and biophysical characterisation
4. Structural determination and functional correlation

17.2.1 Proteomic Analysis

The whole machinery of a cell (and even acellular living forms) is operated, multiplied and regulated by the proteins. Thus, the holistic, as well as the individual, study of proteins becomes imperative. Proteomics plays a pivotal role in the identification of novel protein targets which are part of pathogen survival

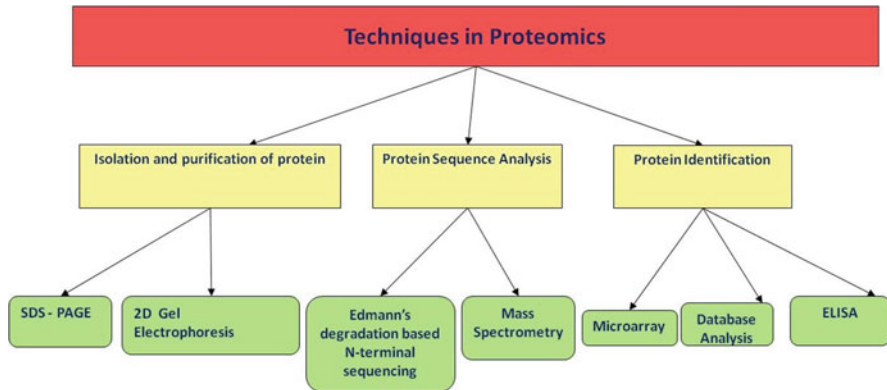


Fig. 17.1 Major techniques used in proteomic approach

strategies, defence responses by the host and subsequently, the host-pathogen interactions. The differential proteomics investigations output evolved from a large number of proteins laid the foundation of new hypotheses and verification of their functions. Intracellular bacteria have evolved with different mechanisms to interfere with host defence system and are successful as a causative agent of many infectious diseases in human. Particularly, intracellular vacuoles are used as an essential niche by these pathogens to overtake the cellular functions facilitating their replication and survival. *M. tb*, the causative agent of tuberculosis in human, has not only been able to alter its intraphagosomal fate by blocking phagosome maturation but has also devised strategies to withhold the actions by immune cells and to survive for the long term in the host successfully. Proteomics based on mass spectrometry (MS) and two-dimensional gel electrophoresis (2D-PAGE) followed by western transfer and assisted by N-terminal sequencing (Edman degradation) helps in the identification and quantitative analysis of complex protein mixtures (Fig. 17.1). It is increasingly employed to investigate host-pathogen interactions (Fig. 17.1).

M. tb, the causative agent of tuberculosis, has got much attention from the research community in the near past. Still, it remains one of the leading causes of death and sufferings caused by an infective disease. Today, we can decipher the nucleotide sequence of a prokaryotic genome within hours. However, based on genomic sequence information, it remains unpredictable to characterise the functional properties. Till now, major efforts have focused on the features of the genomic organisation of the tuberculosis pathogen. The genome of more than 10,000 different *M. tb* strains with varying genotypes and phenotypes has been studied. However, to decipher the causes behind drug resistance and pathogenicity, the application of the whole genome sequencing technology and comparative analysis proves to be limited. The majority of the point mutations that distinguish groups of strains have been

found in the promoter regions of the genes and/or regions encoding proteins with a hypothetical function and playing an unknown role in the physiology of mycobacteria. In this context, a functional analysis of the information deciphered in the pathogen genome performed using proteomic testing, including quantitative proteomics, becomes relevant.

The features of the organisation of the cell wall, which is resistant to environmental factors, acids and alkalis, make *M. tb* a rather complex target for proteomic analysis. This, in its turn, requires the development of unique conditions for protein extraction. The implemented protocols of proteomic analysis of *M. tb* should also be sufficiently effective, taking into account the complexity of accumulation of a large bacterial mass due to the prolonged culture growth.

17.2.1.1 Major Gains Using Proteomics in Case of *M. tb*

Proteomic profiles of virulent H37Rv strains and avirulent strains of *M. tb* are being used to identify potential candidate for vaccine. Proteomic characterisation of H37Rv suggests a change in DosR regulon protein in hypoxic condition. Comparative proteomic analysis of the proteins of a latent H37Rv strains at the exponential, logarithmic and stationary growth phases was done using the technique of site-specific labelling of cysteine residues (isotope-coded affinity tags, CAT) based on covalent labelling of the cysteine residue in the polypeptide chain by chemically identical but isotopically different reagents. The study showed 193 and 241 proteins presents in exponential and stationary phase, respectively, mostly associated with energy metabolism and protein degradation.

Definition of Mycobacterium tuberculosis culture filtrate proteins by two-dimensional polyacrylamide gel electrophoresis, N-terminal amino acid sequencing, and electrospray mass spectrometry by M.G. Sonnenberg and J.T. Belisle analyse 32 proteins.	1997
Comparative proteome analysis of Mycobacterium tuberculosis and Bovis BCG strains toward functional genomics of microbial pathogens by P.R. Jungblut <i>et al.</i> analyse 107 proteins.	1999
Toward the proteome of Mycobacterium tuberculosis by I. Rosenkrands <i>et al.</i> analyse 167 proteins.	2000
Comprehensive proteomic profiling of the membrane constituents of a Mycobacterium tuberculosis strain by S. Gu <i>et al.</i> analyse 739 proteins.	2003
Complementary analysis of the Mycobacterium tuberculosis proteome by two-dimensional electrophoresis and isotope-coded affinity tag technology by F. Schmidt <i>et al.</i> analyse 361 proteins.	2004
Mycobacterium tuberculosis functional network analysis by global subcellular protein profiling K.G. Mawuenyega <i>et al.</i> analyse 1044 proteins.	2005
Using a label-free proteomics method to identify differentially abundant proteins in closely related hypo- and hypervirulent clinical Mycobacterium tuberculosis Beijing isolates by G.A. de Souza <i>et al.</i> analyse 1668 proteins.	2010
Comparison of membrane proteins of Mycobacterium tuberculosis H37Rv and H37Ra strains by H. Mäden <i>et al.</i> analyse 1578 proteins.	2011
Characterization of the Mycobacterium tuberculosis proteome by liquid chromatography mass spectrometry-based proteomics a comprehensive resource for tuberculosis research by C. Bell <i>et al.</i> analyse 1051 proteins.	2011
Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry D.S. Kelkar <i>et al.</i> analyse 3176 proteins.	2011
The Mtb Proteome Library: A resource of assays to quantify the complete proteome of Mycobacterium tuberculosis O.T. Schubert <i>et al.</i> analyse 3894 proteins.	2013
Disclosure of selective advantages in the "modern" sublineage of the Mycobacterium tuberculosis Beijing genotype by J. de Keijzer <i>et al.</i> analyse 2392 proteins.	2014
Quantitative proteomic analysis of M. tuberculosis cluster Beijing B0/W148 strains by J. Bespyatykh <i>et al.</i> analyse 1868 proteins.	2015

Similarly, proteomic characterisation of Beijing strain shows its association with drug resistance and highly virulent nature. The comparative analysis with H37Rv shows the protein responsible for virulence factors, i.e. Rv0129c, Rv0831c, Rv1096, Rv3117 and Rv3804c, was higher in the Beijing strain than in H37Rv. Proteins Hsp65 (Rv0440), Pst1 (Rv0934) and Rv1886c are low in Beijing strain which helps to avoid host immune response. Furthermore, proteins of the efflux pump Rv0341, Rv2688c and Rv3728 were found only in the Beijing strains. Nowadays, the post-translational modifications (PTMs) of *M. tb* proteins and their implications are also identified by proteomic analysis like mannosylation decreases the virulence of *M. tb*. Mostly PTMs are responsible for the regulation of enzymatic activity, interaction with other molecules and lifetime of cellular proteins. Similarly, *M. tb* antigens are found in surface heparin-binding haemagglutinin that is used for the design of a new vaccine.

17.2.1.2 For the Proteomic Analysis, the Two-Way Approach Is Followed, Which Is Represented in the Flowchart (Figs. 17.2 and 17.3)

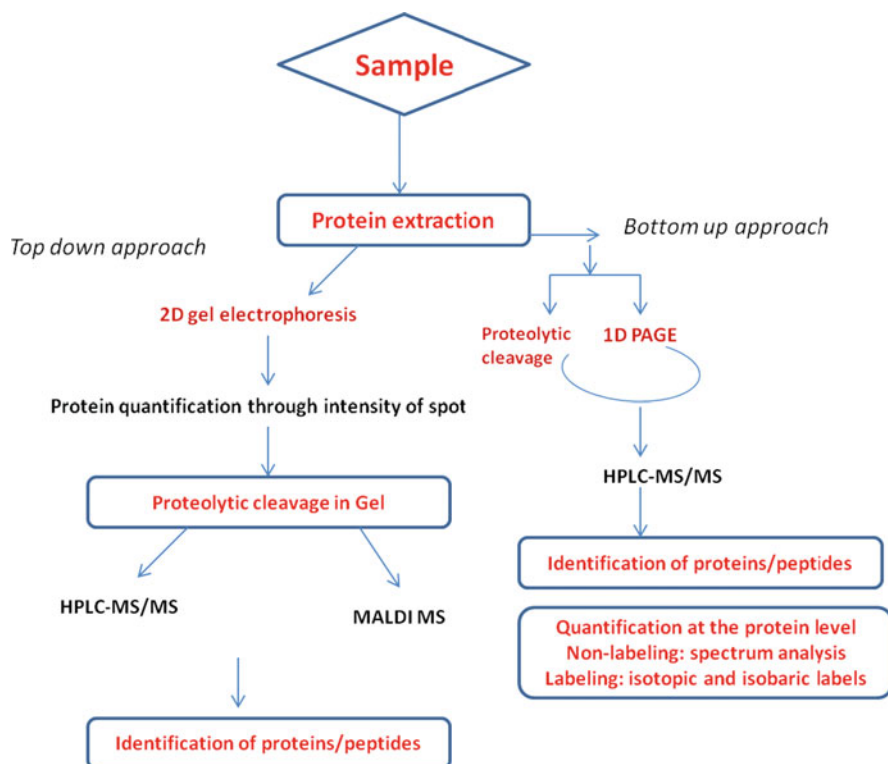


Fig. 17.2 Depiction of major techniques used in the proteomic analysis

17.2.1.3 Major Techniques Used in the Proteomic Analysis

2D Gel Electrophoresis

This technique is most currently used and done before MS and bioinformatic analysis. It separates proteins according to their isoelectric points and molecular weights, and the major advantage of this technique is to separate differentially post-translationally modified forms of the same protein.

Mass Spectroscopy

The principle of MS analysis involves the conversion of the subject molecules to either cations or anions in the ion source, separation according to their mass/charge (m/z) ratios in the mass analyser and subsequent detection. Several configurations of mass spectrometers that combine ES and MALDI with a variety of mass analysers (linear quadrupole mass filter [Q], time-of-flight [ToF], quadrupole ion trap and Fourier transform ion cyclotron resonance [FTICR] instrument) are routinely used.

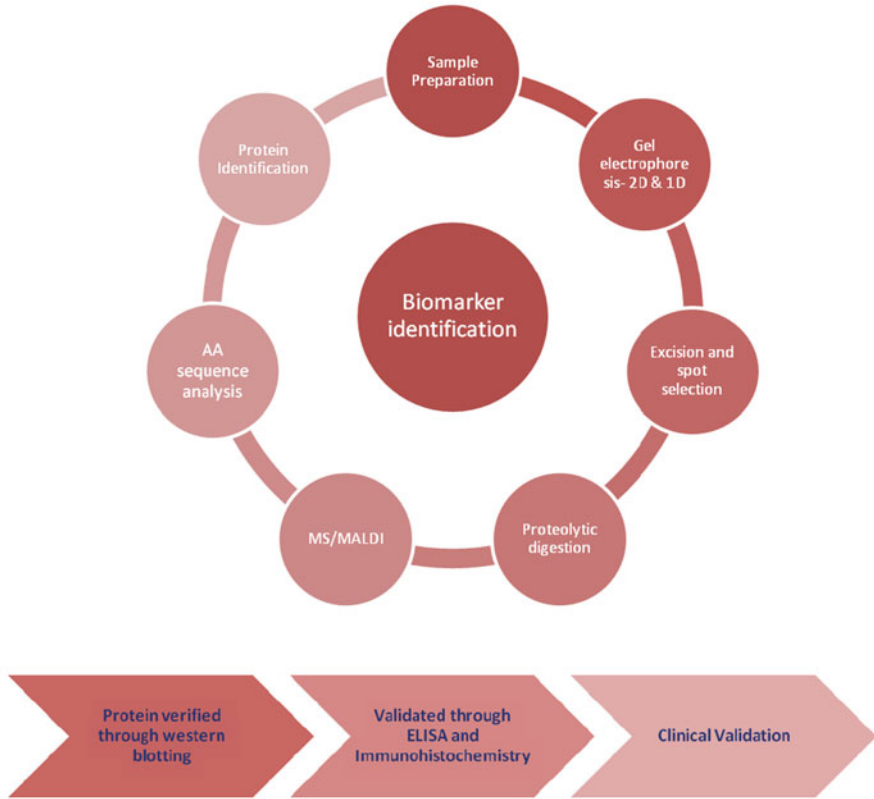


Fig. 17.3 A proteomic approach for biomarker identification and validation

Edman Sequencing

Edman degradation is used to identify the sequence of a protein through labelling and cleaving the peptide without damaging the whole protein in hydrolytic condition. Phenyl isothiocyanate forms phenyl thiocarbamoyl derivative with the N-terminal in less acidic condition form cyclic derivative of PTH. This method can be repeated for the rest of the residues, separating one residue at a time.

ELISA (Enzyme-Linked Immunosorbent Assay)

In this technique antigen-antibody is used, it includes an enzyme-labelled antigen or antibody, and enzyme activity is measured calorimetrically. The enzyme activity is measured using a substrate that changes colour when modified by the enzyme. The light absorption of the product formed after substrate addition is measured and converted to numeric values. Depending on the antigen-antibody combination, the assay is called a direct ELISA, indirect ELISA, sandwich ELISA, competitive ELISA, etc.

Microarray

Microarray is a high-throughput technique used to identify and quantify a large number of proteins in a short time span. In the case of *M. tb*, it is effectively used for proteomic analysis of cell lysate, protein constituent of culture media as well as analysis of pleural fluid and other samples of tuberculosis patients.

17.2.2 Bioinformatic Analysis of Mycobacterial Proteome

Bioinformatic analysis of *M. tb* proteome has explored the host-pathogen interactions and immunomodulation by mycobacterial infection, protein-drug interactions, epitope-driven vaccine candidate and protein-protein interactions. The complex host-pathogen interactions should be quantified consistently for precision and personalised health care [23]. The structural analysis of the protein involves the determination of primary, secondary, tertiary and quaternary structure. Till now, 4983 crystal structures of *M. tb* proteins have been submitted in the Protein Data Bank repository which includes proteins (apoprotein, protein-ligand complex, DNA binding proteins, RNA binding proteins, small peptides and post-translationally modified proteins), nucleic acid and carbohydrates. The structural analysis explains the mechanism of action at the molecular levels and provides leads for the development of drug/vaccine. Proteomic analysis can be divided into the following sections.

17.2.2.1 Database/Tools of Virulence Factors

Virulence factors (VF) are the proteins which are involved in the host-pathogen interactions, the progression of disease and survival inside the host macrophages. There are several online and standalone tools, and databases of VF especially developed for *M. tb* or bacteria or to different disease-causing pathogens. The Pathosystems Resource Integration Centre (PATRIC) is a repository of integrated omics datasets (genome, transcriptome, protein-protein interactions, the 3D structure of proteins and sequenced data) [26]. The virulence factor database (VFDB) contains the whole genome, sequence, structural and functional data which emphasise on the common and species-or strain-specific VFs. VFDB pipeline is based on comparative pathogenomics and VF analysis. The newly developed VFDB 2019 version has additional pathogenic genera *Francisella* and *Klebsiella* [27]. There are other tools which are based on machine learning approaches like VCIMPred, VirulentPred, EffectiveT3 and T4EffPred for the recognition of VFs. VRprofile is a bacterial genome-based server which works on backend database named MobilomeDB that comprises gene cluster loci of bacterial mobile genetic elements which include type III, IV, VI and/or VII secretion apparatuses (T3SSs, T4SSs, T6SSs and/or T7SS), integrative conjugative elements (ICEs), prophages, class I integrons, insertion sequence (IS) elements, pathogenicity islands (PAIs) and antibiotic resistance islands (GIs) (ARIs) [28].

17.2.2.2 Interactome Analysis of *M. tb*

TargetTB, MycoPrint, SinCRE-structural interactome computational resource, CHOPIN, *Mtb*-HID (*M. tb*-Human Protein-Protein Interaction Map), Prediction of Pathogenic Proteins in Metagenomic Datasets (or MP3) and PHI-base (pathogen-host interactions base) are the tools and databases which explore the host-pathogen interactions and protein-protein interactions based on genome, proteome, sequence, structural and functional annotations. STRING database has been utilised to decipher the protein-protein interactions of *M. tb* H37Rv. It helped to unravel the possible pathways which lead to drug resistance [29]. TargetTB involves the structural analysis of genome, reactome, interactome, sequence, experimentally validated phenotypic essential data and assessment of the drug on a structural basis by the novel algorithm [30]. MycoPrint and VirulentPred are based on support vector machine (SVM) for the exploration of the interactome of *M. tb* [31]. MP3 is based on SVM and hidden Markov model for the prediction of pathogenic proteins [32]. Human-*M. tb* interactions can be analysed by *M. tb*-HID database, which consists of interactions between five *M. tb* strains (H37Rv, H37Ra, ATCC, 35801/TMC 107/Erdman, ATCC 35801 and CAS_NITR204) and the human host. SinCRE interactome has been developed based on protein sequence, domain, functional annotation and tertiary structure which decipher protein-protein and protein-drug interactions and mutation potentially leading to drug resistance [32]. CHOPIN is a web-based interactome which deciphers the drug resistance associated with the mutation based on the structure of the protein (25833954). PHI-base is a database of host-pathogen interactions between several disease-causing agents and host [33]. These host-pathogen, protein-protein and protein-drug interactions which lead to immunomodulation in the host may play an important role in the development of drugs and vaccines.

17.2.2.3 Structural and Functional Analysis of *M. tb* Proteome in Drug Discovery

Structural analysis of *M. tb* proteome involves X-ray & NMR. The functional analysis is involved in the annotation of the protein based on its cellular, biological or molecular function. AgBase-GOanna v.2 is a web-based database which provides information for gene ontology [34]. Drug designing is one of the strategies based on the structural analysis for the development of novel drugs. It involves structure-based and ligand-based drug designing, quantitative structure-activity relationship (QSAR), pharmacophore modelling and virtual high-throughput screening for the scrutinisation of the drug molecule [35]. The potential drug candidates are further analysed by structure-activity relationship (SAR) to determine their specificity, sensitivity and activity [35]. Structural annotation of *M. tb* has been carried out based on its genome and proteome which has shown its impact on the pathogenesis, virulence, drug discovery, new drug identification and structure-based lead design [36].

17.2.3 Biochemical and Biophysical Characterisation

The biophysical and biochemical characterisation lays the essential foundation work for structure and functional studies of novel targets, by helping them to crystallise. For example, in 2018, Yang et al. cloned, expressed and purified the transcription factor EF-TU of *M. tb*, which may be used as a novel drug target. Dynamic light scattering suggested that it is present as monomeric form, and circular dichroism showed it is a well-structured protein. The ITC indicated that it has an intermediate affinity towards GTP and GDP, while ES-MS determined the molecular weight of protein. The structure modelling through docking suggested that they generally bind through H bonds. These experiments helped to understand the chemistry of the protein and their binding properties which also helped to explore the biochemical properties of the protein.

Fluorescence spectrometry helps us in the standardisation of emission/absorption maxima at a particular wavelength by measuring the fluorescence intensity at the particular wavelength. This can also be used for assessing any structural change in the protein as well as any protein-protein or protein-ligand binding interactions, again helpful in drug design.

Western blotting, followed by antibody binding and ELISA, is the most widely used confirmatory technique for biochemical and functional characterisation. The techniques are not only helpful in the quantification of particular protein but are also helpful in drug targeting and diagnostic developments. In drug targeting and diagnostic design, they have immense importance in dosimetry.

17.2.4 Structural Determination and Functional Correlation

Proteins are involved in many biological pathways as catalysing agent/inhibitors/activators/modifiers by interacting with other macromolecules. Structure determination is a process by which the three-dimensional atomic coordinates and interaction within and other macromolecule are observed using X-ray crystallography, NMR spectroscopy, cryo-electron microscopy and molecular modelling. These noncovalent interactions help to stabilise the 3D structure of the molecule. The specificity and affinity of these interactions are exclusive to biological functions and facilitate many chemical and physical processes. X-ray crystallography is a method to determine the interatomic spacing of most crystalline solids by putting them as a diffraction gradient of X-rays of wavelength 1 Å in order to produce a high-quality structure, while in NMR the applied magnetic field orients atomic nuclei within or against the field. The nuclei absorb EM radiation to fill this energy gap which determines the composition of mixture or product formed during the reaction and also determines the number of hydrogens attached to each carbon. Cryo-EM is a revolutionary technique taking over X-ray crystallography to determine the structure by exposing the flash freeze sample in an electron beam in order to reconstruct the structure of protein and protein complex, and it helps to visualise the protein which is not able to crystallise. This can be well exemplified by the study of

Lin et al. 2017 who determined the structure of *M. tb* RNA polymerase at 3.8 Å through X-ray crystallography which is the target of first-line drug rifampicin which revealed that it inhibits the extension of RNA through steric occlusion mechanism. Also, non-related RIF compound inhibits the RNA polymerase with no cross-reaction with rifampicin, so if administered together, it will inhibit the growth of *M. tb* effectively (PDB ID 5UHA). Likewise, CryoEM structure of *M. tb* 50S and 70S ribosomal subunit determined by Yang et al. in 2019 suggested that the inter-bridge in 70S helps to understand the structural basis of translation in *M. tb* which led to the development of new drugs (PDB ID 5V93).

17.3 Differential Proteomic Expression Profiles of H₃₇Rv, H₃₇Ra and BCG and Other Clinical Isolates

The accessibility of genome sequences [37] has facilitated genome-wide comparisons of different mycobacterial species to identify gene mutations, deletions or insertions correlated with its virulence and pathogenic characteristics [38]. Proteomic analyses basically counterpart the genomic data in presenting whichever genes are really expressing and reflect the functional status of the cell under different environment conditions [39]. The availability can expect advancement in the development of novel TB therapeutic measures and vaccine candidates of large-scale differential mycobacterial proteome data.

Proteomic analyses of different mycobacterium species and strains (*M. tb* H₃₇Rv, H₃₇Ra, BCG) have highlighted the importance of varied gene expression profiles of several proteins involved in survival strategies of the pathogen, emergence of drug resistance, host-pathogen interaction, etc. [40]. Proteome exploration of cellular proteins of *M. tb* and BCG strains demonstrates 13 proteins unique to *M. tb* H₃₇Rv and 8 to BCG. These differences in the protein composition between attenuated and virulent strains of *M. tb* are supportive for the development of novel vaccine candidates and therapeutics [41].

Singh et al. compared proteomes of 12 different pathogenic mycobacterial strains using Insilco tools to investigate the virulence factors of the species and compared with 241 experimentally validated virulence factors of *M. tb* H₃₇Rv. True, opportunistic and non-pathogenic strains have been found to share 66%, 52% and 34% identity, respectively, with *M. tb* virulent proteins. Conserved nature of virulent proteins among the genus *Mycobacterium* points towards their co-option and evolution. Insilco comparative analysis of *M. tb* with different opportunistic pathogens has shown variable expression and sequence similarities. Proteins belonging to phospholipase, transferase and ESX were observed to be less similar with *Mycobacterium indicus pranii* (MIP) proteome. Indeed four *M. tb* proteins from phospholipase family were shown to present in all other pathogenic mycobacterial members but absent in *M. leprae*. Unique conservation of 14 ESX-related proteins was found in *M. bovis* and *M. marinum*. Homologues of *M. tb* chorismate pyruvate lyase, Fad22, Mmp117 and lipoprotein Lppx are present in pathogenic mycobacterial strains, thereby highlighting their significance in the virulence and pathogenesis. Two

M. tb glycosyl transferases were conserved among *M. tb* H₃₇Rv, *M. bovis*, *M. leprae* and *M. marinum*. One methyltransferase of *M. tb* is highly similar to that of *M. bovis* and *M. leprae* [42]. The conservation of these transferases emphasises the presumed importance of post-translational modifications in mycobacterial virulence mechanism [43].

Another class of *M. tb* proteins called PE/PPE/PE_PGRS family is differentially regulated in different strains of mycobacteria. Around 26 PE and 58 PPE proteins of *M. tb* H₃₇Rv were found to be similar with 7 PE and 21 PPE proteins of MIP, respectively, pointing their importance in the modulation of host immune responses to favour the bacterial survival [42].

In a global protein-protein interaction study by proteome microarray in a yeast expression system, 14 different serum *M. tb* proteins were found to distinguish active TB patients from recovered individuals. These proteins can, therefore, be used as biomarkers for assessing the treatment outcomes [44]. Besides, proteome microarray revealed the likely regulation of *M. tb* rhamnose pathway by c-di-GMP (ubiquitous bacterial secondary messenger) [45] and Ser/Thr kinase PknG. This is associated with cell wall synthesis [46]. It shows that *M. tb* proteome microarray study can be applied to recognise novel molecular targets to combat TB.

The secreted proteins of *M. tb* that are exported intracellularly to host cells were thought to participate in the phagosomal remodelling and bacterial survival inside phagosomal compartments [47]. A study was conducted to identify expression profiles of culture supernatant proteins of *M. tb* virulent H₃₇Rv and avirulent H₃₇Ra strains by 2D gel electrophoresis. Protein expression of Rv2346c, Rv2347c, Rv1038c and Rv3620c has been evident in *M. tb* virulent strain but absent in avirulent bacterium. The location of these protein-coding genes traced in the corresponding area in *M. tb* H₃₇Ra and multiple mutations was found to be associated with their different expression profiles. The 59th codon CAG coding for glutamine in virulent strain was replaced with a stop codon in H₃₇Ra avirulent strain. This difference in the expression profiles is most likely associated with the attenuating characteristic of avirulent bacterium [40].

Quantitative proteomics studies have highlighted substantial distinctive proteome expression profiles of *M. tb* H₃₇Rv and *M. bovis* BCG. The majority of differential expression has been designated particularly to pathways involved in lipid biosynthesis [41, 48] and different growth phases accompanied by nutrient starvation [49, 50]. Furthermore, another study demonstrated the varied proteome profiles of seven clinically significant mycobacterial species from MTBC complex including four *M. tb* strains (H₃₇RV, LAM, Beijing, CAS), *M. avium*, *M. bovis* and *M. bovis* BCG using LC-MS/MS technique. The objective was to identify relevant phenotypic disparities in different pathogenic strains in terms of immune response generation, virulence and transmission. A total of 3788 unique *M. tb* proteins out of 4023 theoretical proteomes have been identified. Each of the MTBC members has been identified with an average of 3290 unique proteins, which represent around 82% of their total theoretical proteome. *M. avium* is represented by 4250 unique proteins that comprised 80% of its theoretical proteome. Although all the classes of proteins were found to be expressed in all strains, the significant quantitative difference was

reported between different strains. Relative expression differences in virulence-related proteins have been shown among different strains contributing to bacterial fitness. A total of 989 proteins were similar in four *M. tb* strains, *M. bovis* and *M. bovis BCG*, but they do not share similarity with *M. avium*. Expression pattern of 168 proteins was uniquely present in *M. tb* strains only and absent in either of *M. bovis BCG*, *M. bovis* or *M. avium*. These unique proteins can, therefore, be allocated as virulence factors of *M. tb* that can be explored further. These unique 168 proteins were expressed under different conditions such as starvation, macrophage infection, hypoxia and acidic models that prove the plausible roles of these proteins in conferring an advantage to the bacterium in vivo [51].

PyrB, PyrC, CarA and CarB proteins are present in *M. tb* pyrimidine biosynthetic operon involved in drug resistance. The Beijing strain of MTBC has been shown to express the highest levels of CarA, PyrB and PyrC. Rv0966 is overexpressed in Latin American Mediterranean (*M. tb* LAM strain) only. Rv2108, Rv2136c, Rv1002c and Rv2703 are reported to be expressed at higher levels in the Beijing strain. Central Asian strain lineage (CAS strain) of *M. tb* when compared to other strains specifically upregulates Rv1818c expression. Rv0096, Rv2108, Rv2136c, Rv1002c and Rv2703 have been hypothesised to modulate host functions, and their expression level was enhanced in pathogenic *M. tb* strains only. Rv2136c is a well-known virulence factor of MTBC encoding undecaprenyl pyrophosphate phosphatase involved in lipid biosynthesis [51, 52]. Rv0833 encoding PE_PGRS33 and Rv3340 encoding O-acetyl homoserine sulphydrylase MetC are involved in the growth of *M. tb* and appeared to be highly abundant in *M. tb* LAM strain, *M. tb* H₃₇Rv and *M. bovis* in comparison to BCG and *M. avium*. Only *M. tb* LAM and *M. avium* were shown to have upregulated expression of Rv3621c encoded PPE65. An important toxin protein VapC2 encoded by Rv0301 has reported having the highest expression in the Beijing strain [53].

Within host macrophage, *M. tb* generally depends on the intracellular setting for the majority of the carbon and iron sources [54, 55]. Thus, in terms of adaptation, Rv1346 (Acyl-CoA dehydrogenase MbtN) is profusely expressed in *M. tb* Beijing strain. Rv3709c encoded aspartate kinase (Ask) was comparatively more abundant in *M. tb* H₃₇Rv and LAM but almost absent in *M. bovis* and *M. avium*. Relatively higher abundance of MetC and aspartate kinase in *M. tb* LAM strain provides the capacity to synthesise essential amino acids for the bacterium, thereby selectively favouring the bacterial virulence [51]. Rv1346 that encodes for acyl-CoA dehydrogenase, which is an enzyme required for the production of mycobactin (an essential molecule for iron acquisition in infected macrophages), is another important virulence factor for *M. tb* infection [56].

However, another systematic proteomic profiling of two *M. tb* strains H₃₇Ra and H₃₇Rv along with two clinical isolates BND-433 and JAL-2287 (belonging to CAS lineage) revealed significant insights into the differential protein expression patterns contributing to the differences in drug resistance and virulence capacity. Out of the total 2161 protein groups identified, which covered 54% of *M. tb* proteome, 257 protein groups were reported to be differentially expressed among different clinical isolates. A total of 13, 12, 13 and 22 proteins groups were found to be specifically

expressed in H₃₇Rv, H₃₇Ra, BND and JAL strains, respectively. The majority of proteins expressed from 2161 *M. tb* proteins were reported to belong cell wall and cell processes (17.2%), intermediary metabolism and respiration (31%) and conserved hypothetical (22.2%), when categorised based on their functional importance [57].

The proteome of JAL strain was found to be significantly distinct in comparison to H₃₇Rv, H₃₇Ra and BND. Cluster gram of identified differentially expressed proteins showed major up- and downregulation of several proteins as compared to other strains. A substantial variation in the regulatory proteins such as transcription factor has been described which has possibly accounted for its intricate regulatory mechanism. Expression profiles of Mce1 and ESX operon proteins in BND and JAL strains, respectively, were very discrete. Proteins expressed from Mce1 operon are expressed in lesser amounts in BND strain [57]. Abrogation of an ESX protein ESAT-6 contributed to the diminished virulence capability of JAL but had minimal effects on other strains. *M. tb* H₃₇Ra completely lacks ESAT-6 and is considered to be an avirulent strain in nature [58]. An interesting finding pointed the lower levels of Rv2780 encoding L-alanine dehydrogenase in JAL strain, which was the first identified antigen known to be absent in *M. bovis* BCG vaccine strain [41, 59].

Ninety of the 257 differentially expressed proteins have been identified as enzymes participating in 29 different metabolic cascades. Among those, five belong to membrane metabolism, and the other four belongs to redox metabolism. Both these differentially expressed protein groups are upregulated in JAL strain specifically. Among the 90 differential proteins, 12 have been specified to account in lipid metabolism performing the particular function “beta-oxidation of fatty acids” in all strains (*M. tb* H37RV, H37Ra, JAL and BND). One and six proteins from this group have been proved to be downregulated in *M. tb* H₃₇RV and *M. tb* JAL clinical isolate, respectively. Moreover, this downregulation has not affected the overall lipid metabolism due to the overexpression of other redundant protein in these strains [57].

Altogether, these studies have pointed out a significant role of differential proteome profiles of different mycobacterium strains, which eventually affect the virulence and pathobiology of the pathogen during infection. Study of the distinct protein expression profiles of different strains will aid in better understanding of the biology of the pathogen.

17.4 Proteomics Studies Delineating the Differences Between Drug-Resistant and Sensitive Strains of *M. tb*

Tuberculosis caused by *M. tb* is an important public health concern, and the spike in drug-resistant TB cases has aggravated the issue. As per 2019 WHO report MDR/RR-TB cases accounted for 3.4% and 18% of new and previously treated cases, respectively, and the success rates of treatment of MDR and XDR-TB is just 50% and 34%, respectively [60]. So, to achieve the END TB strategy, a better understanding of the pathogenesis as well as the resistance mechanism of MDR-TB is

needed. Mutations in KatG, RpoB, GyrA, GyrB, InhA, PncA, AhpC, EmbB, Rrs, GidB and RpsL account for 36–95% of *M. tb*-resistant strains. The remaining 5–64% can be accounted for by some other resistance mechanisms like overexpression of porins, efflux pumps [61] or proteins neutralising drugs activity, making proteomics study necessary. Genomics and transcriptomics have provided great insight into TB pathogenesis, but the actual state of the cell can be more informative when it comes to identifying metabolic and physiological characteristics responsible for infection and drug resistance. So proteomic profiling is the new means of unravelling unique features of *M. tb* strains leading to a different degree of virulence and drug sensitivity.

Proteomics of drug-resistant *M. tb* has garnered much attention in the last decade with numerous studies giving valuable information about the different drug resistance markers and mechanisms adopted by the bug. Proteomics has been used for exploring the metabolic pathways involved in the action of different anti-TB drugs.

A study comparing proteomes of isoniazid-resistant and sensitive strains revealed five upregulated membrane proteins in resistant strains which are electron transfer flavoprotein FixB (Rv3028c), oxidoreductase (Rv2971), Wag31 (Rv2145c), OpcA (Rv1446c) and RegX3 (Rv0491) [62]. RegX3 belongs to a two-component response regulator enabling mycobacteria to adapt to stress posed by antibiotics. These results did not corroborate with earlier known mechanism of INH resistance, namely, KatG mutations. This points towards the likelihood of these proteins to be exploited as hits for novel therapeutic agents in future.

Another study involving RIF- and INH-resistant isolates found 27 consistently overexpressed proteins, out of which most prominent were Wag31 (Rv2145c), GarA (Rv1827), Rv1437 and Rv2970c [63]. Wag31 presence in both studies makes it an important candidate for further studies. It is earlier known to be a homologue of DivIVA and a substrate for PknA and PknB. GarA (glycogen accumulation regulator protein) plays a role in TCA cycle and metabolism of glutamate and also required for intracellular growth of *M. tb* inside macrophages. It is also observed that rifampicin resistance occurs because of some different mechanisms other than well-characterised RpoB mutation based on the study showing underexpression of four proteins (FabD, Ino1, PPE60 and EsxK) in rifampicin-treated *M. tb* [64]. These proteins play a role in cell wall biosynthesis.

Similarly, another study focussed on aminoglycoside (amikacin and kanamycin)-resistant strains employing 2DE coupled with MALDI-TOF/TOF-MS and bioinformatics. Proteins showing increased intensities in resistant isolates using PDQuest advanced software were identified as ferritin (Rv3841), putative short-chain type dehydrogenase/reductase (Rv0148 and Rv3224), bacterioferritin (Rv1876), elongation factor Tu (Rv0685), ATP synthase subunit alpha (Rv1308), alpha-crystallin/HspX (Rv2031c), proteasome subunit alpha (Rv2109c), trigger factor (Rv2462c), 35 kDa hypothetical protein (Rv2744c), transcriptional regulator MoxR1 (Rv1479), dihydrolipoyl dehydrogenase (Rv0462) and universal stress protein (Rv2005c) [65]. Of these, Rv3841, Rv3224, Rv1876 and Rv0685 play an important role in iron metabolism. Acquisition of iron is known to be important due to its role in

mycobacterial growth, virulence and dormancy. Rv3841 (ferritin) maintains iron homeostasis in mycobacterial cells and makes it recalcitrant to antibiotics [66].

Streptomycin-resistant *M. tb* when compared with susceptible strains showed differential expression of 15 proteins which are Rv0350, Rv0440, Rv3846, Rv1860, Rv1636, Rv3418c, Rv1980c, Rv3248c, Rv2140c, Rv1926c, Rv0896, Rv3804c, Rv0009, Rv0815c and Rv2334 [67]. Likewise, Rv2896c, Rv2032c, Rv1908c, Rv1827, Rv0635 and Rv0036 were found to be overexpressed in multidrug-resistant strains by a group of researchers. Rv1908c gets activated in phagosomes [68] and is also necessary for growth and persistence in mice, guinea pigs and human peripheral blood monocytes. Rv2896c, Rv2032c, Rv1827, Rv0635 and Rv0036 play a role in intracellular survival.

Proteomic profiling of *M. tb* exposed to hypoxia revealed a strong induction of DevR/DosR regulon proteins. Since drug-resistant strains mimic dormant cells, total quantification of proteins involved in stress response may also provide some leads to drug resistance mechanisms [69]. The proteome analysis of dormant *M. tb* compared with reactivated bacteria at different stages of infection revealed the differential and unique expression profiles of around 1871 proteins accounting for 47% of *M. tb* proteomes. The percentage of proteins identified in different stages of dormancy and reactivation was observed to be very less as compared to that of control. The most significant fluctuations in the expression profiles belong to the proteins involved in the metabolism of cells. Therefore, the proteins that were found to be differentially or uniquely overexpressed in the reactivation stages can serve as promising targets for novel therapeutics and vaccine potential [70]. Mycobacterial Dop protein involved in proteasome-dependent degradation has been upregulated in dormant stages. Degradation of proteasome-mediated Pup-Dop system is important in achieving a dormancy state [70].

Thus, proteomics has helped greatly in deciphering *M. tb* responses when exposed to drugs providing a peep inside mechanisms of drug action and resistance. Taking all the studies together, it can be implied that differential expression of proteins mostly involved in lipid metabolism, virulence, detoxification and adaptation, cell wall and cell processes, ATP-binding cassette transporters and proteasome function between drug-resistant and sensitive strains shifts the attention from conventional drug resistance mechanisms to novel systems affecting drug efficacy. With the help of these studies, Rv2031c, Rv3692 and Rv0444c are narrowed down as biomarkers for diagnosis of MDR-TB [68].

17.5 Conclusions

The gain in momentum in proteomics studies over the years proves to be vital for the effective understanding of tuberculosis disease. Advancement in proteomics has made it easier to investigate the pathogen *M. tb* in more depth, giving insight into the different factors and strategies adopted by the bug for establishing successful infection. This chapter illustrates the role of proteomics in the study of native proteins of *M. tb* involved in virulence, host-pathogen interaction,

immunomodulation and drug resistance. Conventional biomolecule separation techniques like chromatography and gel electrophoresis continue to be exploited to explore *M. tb* proteome. Advanced MS-based approaches have further helped in refining our knowledge of *M. tb* pathogenesis. Analysis of differential expression profiles of diverse proteins among various strains of *M. tb* has provided a comprehensive knowledge of the key players of virulence which includes Fad22, chorismate pyruvate lyase, MMp17, ESX-related proteins, PE/PPE/PE-PGRS protein, lipoprotein Lppx, etc. Rv1038c, Rv2346c, Rv2347c and Rv3620c were found to be expressed exclusively in *M. tb* virulent strains. A majority of proteins showing differential expression in H37Ra, H37Rv, BND and JAL strains belong to intermediary metabolism and respiration (31%), cell wall and cell processes (17.2%) and conserved hypothetical groups (22.2%). Likewise, there have been multiple proteomics studies deciphering pathways and markers involved in multidrug and extensive drug resistance. Majorly upregulated proteins in drug-resistant strains includes Rv3028c, Rv2971, Rv2145c, Rv1446c, Rv0491, Rv1827, Rv1437 and Rv2970c, Rv3841, Rv0148, Rv3224, Rv1876, Rv0685, Rv1308, Rv2031c, Rv2109c, Rv2462c, Rv2744c, Rv1479, Rv0462, Rv2005c, Rv0350, Rv0440, Rv3846, Rv1860, Rv1636, Rv3418c, Rv1980c, Rv3248c, Rv2140c, Rv1926c, Rv0896, Rv3804c, Rv0009, Rv0815c and Rv2334, Rv2896c, Rv2032c, Rv1908c, Rv1827, Rv0635 and Rv0036.

To summarise, the current findings pertain to the undoubted significance of differential expression proteins in all arenas of tuberculosis making proteomics studies indispensable for the development of rapid, simple, cost-effective diagnostics, novel therapeutics and efficient vaccine for management of TB.

References

1. Ukponmwan OE, Ruprecht J, Dzoljic M (1986) An analgesic effect of enkephalinase inhibition is modulated by monoamine oxidase-B and REM sleep deprivations. *Naunyn Schmiedeberg's Arch Pharmacol* 332(4):376–379
2. Rachow A, Ivanova O, Wallis R, Charalambous S, Jani I, Bhatt N et al (2019) TB sequel: incidence, pathogenesis and risk factors of long-term medical and social sequelae of pulmonary TB – a study protocol. *BMC Pulm Med* 19(1):4
3. Dockrell HM, Smith SG (2017) What have we learnt about BCG vaccination in the last 20 years? *Front Immunol* 8:1134
4. Podany AT, Swindells S (2016) Current strategies to treat tuberculosis. *F1000Res*:5
5. Tiberi S, Scardigli A, Centis R, D'Ambrosio L, Munoz-Torrico M, Salazar-Lezama MA et al (2017) Classifying new anti-tuberculosis drugs: rationale and future perspectives. *Int J Infect Dis* 56:181–184
6. Migliori GB, Tiberi S, Zumla A, Petersen E, Chakaya JM, Wejse C et al (2020) MDR/XDR-TB management of patients and contacts: challenges facing the new decade. The 2020 clinical update by the Global Tuberculosis Network. *Int J Infect Dis* 92S:S15–S25
7. Patterson B, Wood R (2019) Is cough really necessary for TB transmission? *Tuberculosis (Edinb)* 117:31–35
8. Chai Q, Wang L, Liu CH, Ge B (2020) New insights into the evasion of host innate immunity by *Mycobacterium tuberculosis*. *Cell Mol Immunol* 17(9):901–913

9. Bradfute SB, Castillo EF, Arko-Mensah J, Chauhan S, Jiang S, Mandell M et al (2013) Autophagy as an immune effector against tuberculosis. *Curr Opin Microbiol* 16(3):355–365
10. Flannagan RS, Cosio G, Grinstein S (2009) Antimicrobial mechanisms of phagocytes and bacterial evasion strategies. *Nat Rev Microbiol* 7(5):355–366
11. Weiss G, Schaible UE (2015) Macrophage defense mechanisms against intracellular bacteria. *Immunol Rev* 264(1):182–203
12. Ehlers S, Schaible UE (2012) The granuloma in tuberculosis: dynamics of a host-pathogen collusion. *Front Immunol* 3:411
13. Qiang L, Wang J, Zhang Y, Ge P, Chai Q, Li B et al (2019) *Mycobacterium tuberculosis* Mce2E suppresses the macrophage innate immune response and promotes epithelial cell proliferation. *Cell Mol Immunol* 16(4):380–391
14. Su H, Zhu S, Zhu L, Kong C, Huang Q, Zhang Z et al (2017) *Mycobacterium tuberculosis* latent antigen Rv2029c from the multistage DNA vaccine A39 drives TH1 responses via TLR-mediated macrophage activation. *Front Microbiol* 8:2266
15. Peng H, Wang X, Barnes PF, Tang H, Townsend JC, Samten B (2011) The *Mycobacterium tuberculosis* early secreted antigenic target of 6 kDa inhibits T cell interferon-gamma production through the p38 mitogen-activated protein kinase pathway. *J Biol Chem* 286(27):24508–24518
16. Wang X, Barnes PF, Huang F, Alvarez IB, Neuenschwander PF, Sherman DR et al (2012) Early secreted antigenic target of 6-kDa protein of *Mycobacterium tuberculosis* primes dendritic cells to stimulate Th17 and inhibit Th1 immune responses. *J Immunol* 189(6):3092–3103
17. Jung BG, Wang X, Yi N, Ma J, Turner J, Samten B (2017) Early secreted antigenic target of 6-kDa of *Mycobacterium tuberculosis* stimulates IL-6 production by macrophages through activation of STAT3. *Sci Rep* 7:40984
18. Refai A, Gritli S, Barbouche MR, Essafi M (2018) *Mycobacterium tuberculosis* virulent factor ESAT-6 drives macrophage differentiation toward the pro-inflammatory M1 phenotype and subsequently switches it to the anti-inflammatory M2 phenotype. *Front Cell Infect Microbiol* 8:327
19. Lin J, Jiang Y, Liu D, Dai X, Wang M, Dai Y (2020) Early secreted antigenic target of 6-kDa of *Mycobacterium tuberculosis* induces transition of macrophages into epithelioid macrophages by downregulating iNOS/NO-mediated H3K27 trimethylation in macrophages. *Mol Immunol* 117:189–200
20. Yaseen I, Kaur P, Nandicoori VK, Khosla S (2015) Mycobacteria modulate host epigenetic machinery by Rv1988 methylation of a non-tail arginine of histone H3. *Nat Commun* 6:8922
21. Schubert OT, Mouritsen J, Ludwig C, Rost HL, Rosenberger G, Arthur PK et al (2013) The Mtb proteome library: a resource of assays to quantify the complete proteome of *Mycobacterium tuberculosis*. *Cell Host Microbe* 13(5):602–612
22. Kunnath-Velayudhan S, Porcelli SA (2013) Recent advances in defining the immunoproteome of *Mycobacterium tuberculosis*. *Front Immunol* 4:335
23. Bespyatykh JA, Shitikov EA, Ilina EN (2017) Proteomics for the investigation of mycobacteria. *Acta Nat* 9(1):15–25
24. Uddin R, Siddiqui QN, Sufian M, Azam SS, Wadood A (2019) Proteome-wide subtractive approach to prioritize a hypothetical protein of XDR-*Mycobacterium tuberculosis* as potential drug target. *Genes Genomics* 41(11):1281–1292
25. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D et al (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393(6685):537–544
26. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL et al (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 42(Database issue):D581–D591
27. Liu B, Zheng D, Jin Q, Chen L, Yang J (2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* 47(D1):D687–DD92

28. Li J, Tai C, Deng Z, Zhong W, He Y, Ou HY (2018) VRprofile: gene-cluster-detection-based profiling of virulence and antibiotic resistance traits encoded within genome sequences of pathogenic bacteria. *Brief Bioinform* 19(4):566–574
29. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M et al (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 45(D1):D362–D368
30. Raman K, Yeturu K, Chandra N (2008) targetTB: a target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. *BMC Syst Biol* 2:109
31. Sharma D, Surolia A (2011) Computational tools to study and understand the intricate biology of mycobacteria. *Tuberculosis (Edinb)* 91(3):273–276
32. Gupta A, Kapil R, Dhakan DB, Sharma VK (2014) MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PLoS One* 9(4):e93907
33. Winnenburg R, Urban M, Beacham A, Baldwin TK, Holland S, Lindeberg M et al (2008) PHI-base update: additions to the pathogen host interaction database. *Nucleic Acids Res* 36 (Database issue):D572–D576
34. McCarthy FM, Wang N, Magee GB, Nanduri B, Lawrence ML, Camon EB et al (2006) AgBase: a functional genomics resource for agriculture. *BMC Genomics* 7:229
35. Macalino SJY, Billones JB, Organo VG, Carrillo MCO (2020) In silico strategies in tuberculosis drug discovery. *Molecules* 25(3):665
36. Chandra N, Sandhya S, Anand P (2014) Structural annotation of the *Mycobacterium tuberculosis* proteome. *Microbiol Spectr* 2(2)
37. Cubillos-Ruiz A, Morales J, Zambrano MM (2008) Analysis of the genetic variation in *Mycobacterium tuberculosis* strains by multiple genome alignments. *BMC Res Notes* 1:110
38. ten Bokum AM, Movahedzadeh F, Frita R, Bancroft GJ, Stoker NG (2008) The case for hypervirulence through gene deletion in *Mycobacterium tuberculosis*. *Trends Microbiol* 16 (9):436–441
39. Rosenkrands I, King A, Weldingh K, Moniatte M, Moertz E, Andersen P (2000) Towards the proteome of *Mycobacterium tuberculosis*. *Electrophoresis* 21(17):3740–3756
40. He XY, Zhuang YH, Zhang XG, Li GL (2003) Comparative proteome analysis of culture supernatant proteins of *Mycobacterium tuberculosis* H37Rv and H37Ra. *Microbes Infect* 5 (10):851–856
41. Jungblut PR, Schaible UE, Mollenkopf HJ, Zimny-Arndt U, Raupach B, Mattow J et al (1999) Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. *Mol Microbiol* 33(6):1103–1117
42. Singh Y, Kohli S, Sowpati DT, Rahman SA, Tyagi AK, Hasnain SE (2014) Gene cooption in mycobacteria and search for virulence attributes: comparative proteomic analyses of *Mycobacterium tuberculosis*, *Mycobacterium indicus pranii* and other mycobacteria. *Int J Med Microbiol* 304(5–6):742–748
43. Wilkinson KA, Newton SM, Stewart GR, Martineau AR, Patel J, Sullivan SM et al (2009) Genetic determination of the effect of post-translational modification on the innate immune response to the 19 kDa lipoprotein of *Mycobacterium tuberculosis*. *BMC Microbiol* 9:93
44. Deng J, Bi L, Zhou L, Guo SJ, Fleming J, Jiang HW et al (2014) *Mycobacterium tuberculosis* proteome microarray for global studies of protein function and immunogenicity. *Cell Rep* 9 (6):2317–2329
45. Romling U, Galperin MY, Gomelsky M (2013) Cyclic di-GMP: the first 25 years of a universal bacterial second messenger. *Microbiol Mol Biol Rev* 77(1):1–52
46. Brennan PJ (2003) Structure, function, and biogenesis of the cell wall of *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* 83(1–3):91–97
47. Berthet FX, Lagranderie M, Gounon P, Laurent-Winter C, Ensergueix D, Chavart P et al (1998) Attenuation of virulence by disruption of the *Mycobacterium tuberculosis* erp gene. *Science* 282(5389):759–762

48. Schmidt F, Donahoe S, Hagens K, Mattow J, Schaible UE, Kaufmann SH et al (2004) Complementary analysis of the *Mycobacterium tuberculosis* proteome by two-dimensional electrophoresis and isotope-coded affinity tag technology. *Mol Cell Proteomics* 3(1):24–42
49. Ang KC, Ibrahim P, Gam LH (2014) Analysis of differentially expressed proteins in late-stationary growth phase of *Mycobacterium tuberculosis* H37Rv. *Biotechnol Appl Biochem* 61(2):153–164
50. Albrethsen J, Agner J, Piersma SR, Hojrup P, Pham TV, Weldingh K et al (2013) Proteomic profiling of *Mycobacterium tuberculosis* identifies nutrient-starvation-responsive toxin-antitoxin systems. *Mol Cell Proteomics* 12(5):1180–1191
51. Peters JS, Calder B, Gonnelli G, Degroeve S, Rajaonarifara E, Mulder N et al (2016) Identification of quantitative proteomic differences between *Mycobacterium tuberculosis* lineages with altered virulence. *Front Microbiol* 7:813
52. Forrellad MA, Klepp LI, Gioffre A, Sabio y Garcia J, Morbidoni HR, de la Paz Santangelo M et al (2013) Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence* 4(1):3–66
53. Sala A, Bordes P, Genevaux P (2014) Multiple toxin-antitoxin systems in *Mycobacterium tuberculosis*. *Toxins (Basel)* 6(3):1002–1020
54. McKinney JD, Honer zu Bentrup K, Munoz-Elias EJ, Miczak A, Chen B, Chan WT et al (2000) Persistence of *Mycobacterium tuberculosis* in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. *Nature* 406(6797):735–738
55. Eisenreich W, Dandekar T, Heesemann J, Goebel W (2010) Carbon metabolism of intracellular bacterial pathogens and possible links to virulence. *Nat Rev Microbiol* 8(6):401–412
56. De Voss JJ, Rutter K, Schroeder BG, Su H, Zhu Y, Barry CE 3rd (2000) The salicylate-derived mycobactin siderophores of *Mycobacterium tuberculosis* are essential for growth in macrophages. *Proc Natl Acad Sci USA* 97(3):1252–1257
57. Jhingan GD, Kumari S, Jamwal SV, Kalam H, Arora D, Jain N et al (2016) Comparative proteomic analyses of avirulent, virulent, and clinical strains of *Mycobacterium tuberculosis* identify strain-specific patterns. *J Biol Chem* 291(27):14257–14273
58. Frigui W, Bottai D, Majlessi L, Monot M, Josselin E, Brodin P et al (2008) Control of *M. tuberculosis* ESAT-6 secretion and specific T cell recognition by PhoP. *PLoS Pathog* 4(2):e33
59. Andersen AB, Andersen P, Ljungqvist L (1992) Structure and function of a 40,000-molecular-weight protein antigen of *Mycobacterium tuberculosis*. *Infect Immun* 60(6):2317–2323
60. Harding E (2020) WHO global progress report on tuberculosis elimination. *Lancet Respir Med* 8(1):19
61. Magnet S, Courvalin P, Lambert T (2001) Resistance-nodulation-cell division-type efflux pump involved in aminoglycoside resistance in *Acinetobacter baumannii* strain BM4454. *Antimicrob Agents Chemother* 45(12):3375–3380
62. Jiang X, Zhang W, Gao F, Huang Y, Lv C, Wang H (2006) Comparison of the proteome of isoniazid-resistant and -susceptible strains of *Mycobacterium tuberculosis*. *Microb Drug Resist* 12(4):231–238
63. Singh A, Gopinath K, Sharma P, Bisht D, Sharma P, Singh N et al (2015) Comparative proteomic analysis of sequential isolates of *Mycobacterium tuberculosis* from a patient with pulmonary tuberculosis turning from drug sensitive to multidrug resistant. *Indian J Med Res* 141(1):27–45
64. Meneguello JE, Arita GS, Silva JVO, Ghiraldi-Lopes LD, Caleffi-Ferracioli KR, Siqueira VLD et al (2020) Insight about cell wall remodulation triggered by rifampicin in *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* 101903:120
65. Sharma D, Kumar B, Lata M, Joshi B, Venkatesan K, Shukla S et al (2015) Comparative proteomic analysis of aminoglycosides resistant and susceptible *Mycobacterium tuberculosis* clinical isolates for exploring potential drug targets. *PLoS One* 10(10):e0139414
66. Pandey R, Rodriguez GM (2012) A ferritin mutant of *Mycobacterium tuberculosis* is highly susceptible to killing by antibiotics and is unable to establish a chronic infection in mice. *Infect Immun* 80(10):3650–3659

67. Sharma D, Bisht D (2017) Secretory proteome analysis of streptomycin-resistant *Mycobacterium tuberculosis* clinical isolates. *SLAS Discov* 22(10):1229–1238
68. Singhal N, Sharma P, Kumar M, Joshi B, Bisht D (2012) Analysis of intracellular expressed proteins of *Mycobacterium tuberculosis* clinical isolates. *Proteome Sci* 10(1):14
69. Sharma D, Bisht D, Khan AU (2018) Potential alternative strategy against drug resistant tuberculosis: a proteomics prospect. *Proteomes* 6(2):26
70. Gopinath V, Raghunandan S, Gomez RL, Jose L, Surendran A, Ramachandran R et al (2015) Profiling the proteome of *Mycobacterium tuberculosis* during dormancy and reactivation. *Mol Cell Proteomics* 14(8):2160–2176



Pathogenesis of *Staphylococcus aureus* and Proteomic Strategies for the Identification of Drug Targets

18

Alaguvel Valliammai, Anthonymuthu Selvaraj, and Shunmugiah Karutha Pandian

Abstract

Staphylococcus aureus is a leading pathogen responsible for mild to severe invasive infections in humans. Especially, methicillin-resistant *Staphylococcus aureus* (MRSA) is prevalent in hospital settings and biomaterial-associated infections. In addition, MRSA is listed as high-priority pathogen in WHO priority pathogen list and occupied the serious threat level in CDC's drug-resistant bacteria report. Persistent *S. aureus* infections are often associated with biofilm formation and resistant to conventional antimicrobial therapy. Inhibiting the surface adherence and virulence of the bacterium is the current alternative approach without affecting growth to reduce the possibility of resistance development. Though numerous antibiofilm agents have been identified, their mode of action remains unclear. Proteomics is the powerful approach to delineate the drug targets of bioactive molecules. Bottom-up strategy-based comparative proteomics is extensively used in the field of disease diagnosis and therapy. Molecular targets of antibiotics and antibiofilm agents active against *S. aureus* have been unveiled using various proteomic approaches and lead to development of drug discovery as well.

Keywords

Staphylococcus aureus · Pathogenesis · Biofilm formation · Proteomics · Drug target discovery

A. Valliammai · A. Selvaraj · S. K. Pandian (✉)
Department of Biotechnology, Alagappa University, Karaikudi, Tamil Nadu, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_18

325

18.1 Introduction

S. aureus is a dangerous bacterium capable of causing deadly invasive infections in humans in addition to mild superficial skin infections. With a plethora of virulence determinants, *S. aureus* is able to adhere to biotic and abiotic surfaces and survive even under adverse host conditions. Especially, *S. aureus* is the predominant cause of biomaterial-associated infections as this pathogen prefers to adhere to the foreign materials inside the body and mostly leads to failure of implanted devices. The major complexity associated with persistent implant infections is formation of biofilm which endows the bacterial cells with the resistance nature [1]. Due to resistance to majority of commonly available antibiotics, the medical community is left with a very few options to treat *S. aureus* infections. Instead of killing the bacteria with antibiotics, inhibition of biofilm formation with antibiofilm agents seems to be a good alternative strategy to fight bacterial infections in the recent times [2]. Apart from finding the potential antibiofilm agents, understanding their mechanism of action is also equally important. Thus, novel drugs with effective mode of action can be synthesized. In addition, toxicity of drug molecules can be ruled out when precise mode of action is known. Proteomic approach gives an in-depth understanding of expression of virulence determinants in *S. aureus* and uncovers the complexity of the virulence machineries involved in pathogenicity. The proteomic approach utilizes various techniques which can be generally classified into two categories, namely, gel-based and gel-free. Various proteomic strategies developed with advancements shed more light on elucidation of drug targets in *S. aureus* antibiotic resistance and contribute to the progression of antistaphylococcal therapy and drug discovery [3]. This chapter elaborates the virulence attributes of *S. aureus* and emphasizes the efficacy of proteomics in drug target identification.

18.2 *S. aureus* Infections in Humans

S. aureus is a human commensal bacterium mostly present in the skin and mucosae. Though various body sites can be colonized by *S. aureus*, the anterior nares of the nose is the frequent and predominant carriage site of *S. aureus* in humans [4]. *S. aureus* colonizes anterior nares of 20–80% of the human population at any stage of life, and 30% of human population is constantly colonized with *S. aureus* [5]. The commensal *S. aureus* turns pathogenic when the individual becomes immune compromised and weak [6]. Various infections caused by *S. aureus* in humans are depicted in Fig. 18.1. *S. aureus* majorly causes skin and soft tissue infections such as abscesses, sores, impetigo, boils, lesions, cellulitis, and folliculitis. Apart from these minor infections, *S. aureus* can also cause life-threatening invasive infections such as bacteremia, pneumonia, osteomyelitis, septic arthritis, otitis media, endocarditis, meningitis, and indwelling device-related infections [7, 8]. In the recent decades, the epidemiology of *S. aureus* gained more global attention because of the high incidence of *S. aureus* in healthcare-associated infections [9].

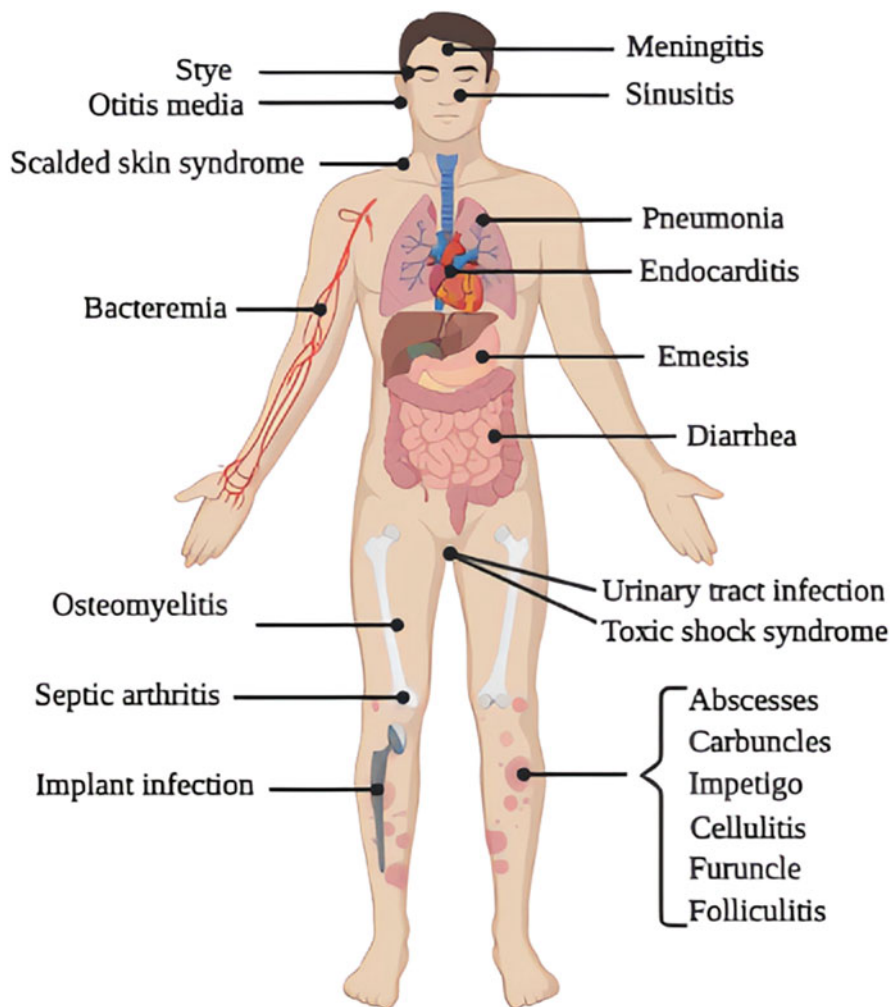


Fig. 18.1 Graphical representation of moderate to severe infections caused by *S. aureus* in humans (Created in [BioRender.com](https://www.biorender.com))

Most notably, *S. aureus* is the predominant pathogen isolated from a variety of implantable medical devices [10]. Extensive usage of implants poses serious problems to the patients by damaging epithelial or mucosal barriers and thereby supports invasion of microorganisms which serve as reservoir of microbial infections [11]. It has been reported that more than 45% of nosocomial infections is caused by means of implanted medical devices. Further research on investigation of microbial community associated with implant infection has revealed *S. aureus* as most dangerous bacterium which can colonize the implanted surface in an irreversible manner [1]. Specifically, *S. aureus* has been frequently encountered in patients

with infective endocarditis and prosthetic device-associated infections [12]. Mortality and morbidity rate of *S. aureus* infections is steadily increasing as prevalence of *S. aureus* became ineradicable [13]. In 2017, the World Health Organization (WHO) released the global priority list of antibiotic-resistant bacteria in which MRSA occupied the high priority [14]. In addition, MRSA was listed as the serious threats in the antibiotic resistance threat report published by the Centers for Disease Control and Prevention (CDC) in 2019 [15].

18.3 Antibiotic Resistance in *S. aureus*: A Global Threat

Interestingly, the world's first antibiotic was discovered from the contaminated plate of *S. aureus*. The story of antibiotics started in the year 1928 when Sir Alexander Fleming, a Scottish physician and microbiologist, accidentally discovered penicillin from the fungus *Penicillium notatum* which contaminated the *S. aureus* plate left opened in his laboratory in Paddington, London [16, 17]. After 12 years from discovery, the pure form of penicillin was made clinically available in the year 1941 which saved the life of numerous soldiers with bacterial pneumonia and meningitis during the Second World War. Due to the ability to cure various bacterial infections, penicillin earned the repute of being called as a “wonder drug” or a “miracle drug.” From then, different classes of novel antibiotics were produced, and commercial availability of many antibiotics happened in the period of 1950–1970 which is referred to as “golden era of antibiotics” [18, 19].

Shockingly, a report on penicillin-resistant *S. aureus* was published in 1940 which was even before the first clinical use of penicillin [20]. Fleming's comment on resistance in his Nobel lecture in the year 1945 was even more surprising. He mentioned that “But I would like to sound one note of warning. Penicillin is to all intents and purposes non-poisonous so there is no need to worry about giving an overdose and poisoning the patient. There may be a danger, though, in under dosage. It is not difficult to make microbes resistant to penicillin in the laboratory by exposing them to concentrations not sufficient to kill them and the same thing has occasionally happened in the body” [21].

Later in 1959, Celbenin (with trade name of methicillin), a penicillinase-resistant penicillin, was launched to fight against penicillin-resistant *S. aureus*. Methicillin was considered to be the end of staphylococcal resistance. However, within a very short span of time, methicillin-resistant *S. aureus* (MRSA) was identified by M. Patricia Jevons from Staphylococcus Reference Laboratory, London, in 1960. *mecA* gene encoding penicillin-binding protein PBP 2A is responsible for methicillin resistance. Over the time, *mecA* gene got spread worldwide and 60–70% of *S. aureus* strains are reported to be methicillin resistant [22]. Vancomycin was approved by the Food and Drug Administration (FDA) in the year 1958 with an aim to treat bacterial strains resistant to methicillin though it was identified earlier in 1953. The first clinical strain of *S. aureus* with reduced susceptibility to vancomycin isolated in Japan in 1996 was named as vancomycin-intermediate *S. aureus* (VISA) and clinical isolate of *S. aureus* with resistance to vancomycin detected in the United States in

2002 was named as vancomycin-resistant *S. aureus* (VRSA) [23, 24]. After the global emergence of VISA and VRSA, the new antibiotic linezolid was approved for clinical use in 2000. Unsurprisingly, linezolid-resistant staphylococci were reported shortly in 2001 [25].

The historical events manifest the ability of *S. aureus* to acquire resistance to a variety of drugs in a short period of time, whereas the discovery of every antibiotic taken a long span of time and immense efforts. From 1970, numerous MRSA strains with multiple drug resistance (MDR) were identified and made MRSA superbug worldwide. Global spread of drug resistance diminished the value of antibiotics in the treatment of bacterial infections [26]. Hence, the MRSA infections are hard to cure with limited efficacy of antibiotics and evolved as serious clinical issue which challenges the clinicians as well as researchers. Evolution of MDR and therapeutic failure of antibiotics led to the post-antibiotic era to overcome severe bacterial infections [27].

18.4 Pathogenesis of *S. aureus*

18.4.1 Repertoire of Virulence Factors

S. aureus is capable of producing plenty of structural and secreted virulence factors involved in pathogenesis. *S. aureus* produces numerous surface proteins, called “microbial surface components recognizing adhesive matrix molecules” (MSCRAMMs), which mediate adherence of bacterial cells to host tissues. MSCRAMMs specifically bind to the major components of host tissue such as collagen, fibronectin, and fibrinogen. MSCRAMMs play a critical role in the initiation of endovascular infections and biomaterial-associated infections. Once *S. aureus* colonized on host tissues or prosthetic surfaces, it is able to survive and persist in several ways. *S. aureus* has various virulence traits to evade the host immune system during establishment of an infection [28]. The major virulence factors produced by *S. aureus* and their role in pathogenesis are presented in Table 18.1.

During the progression of infection, *S. aureus* secretes various enzymes such as elastases, lipases, and proteases to invade and destroy host tissues. These secretory virulence factors help metastasize to the new sites to disseminate the infection. The structural virulence components such as peptidoglycan and lipoteichoic acid play a role in activation of the host immune system and coagulation pathways to produce septic shock. Apart from septic shock, *S. aureus* also produces superantigens that cause toxic shock syndrome and food poisoning [49, 50]. Expression of virulence factors is metabolically expensive and occurs in a highly controlled manner. MSCRAMM adhesion proteins are generally expressed during logarithmic growth phase to facilitate initial adhesion, whereas secretory enzymes and toxins are produced during the stationary phase to progress and disseminate the infection [51].

Table 18.1 Major virulence factors of *S. aureus*

Virulence factor	Biological function	Reference
Adhesins		
Teichoic acid	Highly charged cell wall polymers, play a key role in the first step of biofilm formation	[29]
Intracellular adhesion (Ica)	Synthesize polysaccharide namely poly N-acetyl glucosamine (PIA/PNAG) involved in biofilm formation	[30]
Staphylococcal protein A (Spa)	A surface-anchored structural protein contributes to colonization and immune evasion. Blocks opsonophagocytosis through nonspecific interaction with Fc portion of the immunoglobulin G (IgG)	[31]
Fibronectin-binding proteins (FnbA and FnbB)	Contribute to tissue colonization in various pathological conditions and indwelling medical device-related infections	[32]
Elastin-binding protein (Ebp)	An integral membrane protein mediates adherence of bacterial cells to specific components of extracellular matrix	[33]
Collagen adhesion (Cna)	Facilitates binding of <i>S. aureus</i> to bone matrix	[34]
Clumping factors (ClfA and ClfB)	A cell wall-anchored protein promotes bacterial adhesion to the blood plasma protein fibrinogen and colonization on protein-coated biomaterials	[35]
Autolysin A (AtlA)	A cell surface-associated peptidoglycan hydrolase promotes attachment to polystyrene surfaces and play important role in biofilm development	[36]
Enzymes		
Metalloprotease; aureolysin (Aur)	Belongs to the family of thermolysins, have a role in staphylococcal immune escape by cleaving complement proteins	[37]
Staphopain proteases (SspA, SspB and SspC)	Important immunomodulatory proteins that inhibit phagocytosis and neutrophil recruitment and damage the epithelium and underlying connective tissue. Also involved in biofilm dispersal	[38]
Lipase (Geh/Lip2)	Interfere with the host granulocyte function, and increase survival of the bacteria against the host defense by inactivating bactericidal lipids	[39]
Nuclease (Nuc)	Required for the evasion of neutrophil extracellular traps (NETs) and has a role in the inhibition of biofilm formation	[40]
Catalase (KatA)	An enzyme implicated in oxidative stress resistance and protects intraphagocytic bacteria by destroying hydrogen peroxide produced by the phagocyte	[41]
Hyaluronidase (HysA)	Promotes tissue penetration and disease progression	[42]
Coagulase (Coa)	Activates prothrombin, thereby converting fibrinogen to fibrin and promoting clotting of plasma or blood. Also responsible for abscess formation and persistence in host tissues	[43]

(continued)

Table 18.1 (continued)

Virulence factor	Biological function	Reference
Toxins		
Phenol-soluble modulins (PSMs)	Efficiently lyse white and red blood cells and contribute to the structuring of biofilms and the dissemination of biofilm-associated infections	[44]
Hemolysins (Hld and Hla)	Induces lysis of red blood cells and play an important role in various diseases such as pneumonia, sepsis, septic arthritis, brain abscess, and corneal infections	[45]
Staphylococcal enterotoxins (Sea and Seb)	Enter the bloodstream and circulate through the body, thus allowing the interaction with antigen-presenting cells and T cells that leads to superantigen activity and causes classic food poisoning, nausea, vomiting, and diarrhea without fever	[46]
Panton-Valentine leukocidin (PVL)	A potent pore-forming cytotoxin causes tissue necrosis and selectively disrupts leukocyte membranes, thus leading to enhanced virulence	[47]
Toxic shock syndrome toxin-1 (TSST-1)	A prototype-secreted superantigen binds to class II MHC molecules on antigen-presenting cells and stimulate large populations of T cells leading to an acute toxic shock	[48]

18.4.2 Biofilm Formation

What makes the tiny *S. aureus* to acquire the ability to infect giant humans and even to cause death is just the group behavior. Discovery of bacterial communication otherwise known as quorum sensing not only awakened the global researchers to focus on virulence nature of microorganisms and also unearthed the multi-cellular behavior of microorganisms [52]. Bacteria achieved the eukaryotic lifestyle by adherence based community living in the form of biofilm. Bacterial biofilms are the sessile and highly structured microbial communities which are encased within the self-produced matrix of extracellular polymeric substances (EPS). Biofilm mode of lifestyle enables the bacterium to adhere onto both biotic and abiotic surfaces [53].

S. aureus is a well-known biofilm-producing bacterium and plays a crucial role in hospital-associated infections by forming biofilm on medical devices. Biofilm formation in *S. aureus* is a multistep process. Initial step of biofilm formation is adherence to either biotic or abiotic surfaces using adhesin proteins which is followed by the proliferation of cells to form microcolonies, and then the secretion of EPS induces more cells to form a three-dimensional biofilm. EPS is a hydrated three-dimensional matrix comprising polysaccharides along with molecules such as eDNA, eProteins, and lipids. Once mature biofilm is formed, it becomes a stable microbial community against adverse environmental conditions. Dispersal of biofilm is mediated by the production of matrix-degrading enzymes such as nucleases and proteases [54].

Formation of biofilm provides the adherent stay for the bacteria, thereby making them more virulent than planktonic cells in numerous ways such as resistance to host immune response, altered growth rate, metabolically inactive persister cell

formation, synchronized virulence gene expression, and horizontal gene transfer [55, 56]. Hence, antibiotics are unable to penetrate the slimy EPS matrix of biofilm. In addition, gene encoding antibiotic resistance is highly transferred within biofilm cells, and hence antibiotic-degrading enzymes are overexpressed in biofilm cells, thereby making the whole microbial community antibiotic resistant.

18.4.3 Regulatory Mechanisms of Biofilm Formation and Virulence

The formation of biofilm in *S. aureus* is well organized and tightly controlled as well. A complex network of regulatory molecules controls the expression of biofilm components either positively or negatively [57]. Bacterial cells secrete as well as detect the signaling molecules also called as autoinducing peptides (AIP) which elicit the cascade of biological processes inside a cell. This kind of bacterial communication is called quorum sensing which regulates the expression of virulence traits. In *S. aureus*, quorum sensing is mediated by accessory gene regulatory (*agr*) system which consists of *agrBDCA* operon. AIP-mediated *agr* system regulates the production of an array of structural and secreted virulence factors in *S. aureus*. Agr is a two-component regulatory system controlled by *agr* operon with four genes *agrBDCA* in which *agrD* codes for AIP which is further processed and transported by *agrB* and extracellular AIP is recognized by the receptor protein *agrC* which phosphorylates the cytoplasmic partner *agrA* which further induces the expression of regulatory RNA known as RNAIII as well as induces the expression of *agrBDCA* as feedforward induction. RNAIII inhibits the production of adhesion proteins which are involved in colonization whereas induces the production of matrix-degrading enzymes which are involved in dissemination. Thus, *agr* system acts as the switch between biofilm and planktonic state of bacterial growth depending on the cell density [58, 59].

Apart from *agr* system, various regulators are involved in governing biofilm formation. A major regulatory molecule appears to play a key role is staphylococcal accessory regulator A (*sarA*) which is well reported to positively regulate the biofilm formation. SarA protein has high binding affinity to the promoter region of *ica* operon and induces the production of poly-N-acetylglucosamine (PNAG) also known as PIA which facilitates biofilm formation. Additionally, SarA induces the expression of biofilm-associated adhesive proteins. Further, the stress responsive sigma factor (σ^B) activates *sarA* as well as *ica* operon mediated PIA biosynthesis and supports biofilm formation. On the other hand, MgrA, a well-known member of *sarA* family, impedes biofilm formation by inhibiting the process of autolysis, and it is also involved in activation of *agr* system [57, 60].

18.5 Unveiling the Drug Targets in *S. aureus* Using Proteomic Approaches

18.5.1 Importance of Proteomics in Drug Target Discovery

In recent years, research on proteomics gained more attention because of its ability to extract, separate, analyze, and identify the total proteome. As proteins are functional players arising from genes and being involved in various cellular processes, research on proteome level will enlighten the actual molecular mechanisms of biologically active compounds. In addition, proteomic techniques are highly sensitive and reproducible against a wide range of proteins [61]. The discovery of drug target in any clinically important pathogen is important with a potential health benefits for the welfare of the society. Decoding the principal mechanism of action of a drug and analysis of off-target interactions are essential to explore the therapeutic potential and side effects of the drugs [62]. Proteomics is a robust approach to unveil the mode of action of biologically active molecules against the virulence traits of the pathogens. In addition, proteomics can be exploited to study the quantification of protein abundance, interaction of proteins with other biomacromolecules, and post-translational modifications [63]. The study of proteomics is commonly categorized into two, namely, bottom-up and top-down approaches. The top-down proteomic approach is used to analyze the complex proteins in the intact native state, whereas in the bottom-up approach, proteins are fragmented to peptides prior to analysis and identification. The bottom-up strategy is widely used in the field of health and medicine due to its sensitivity and reliability [64].

18.5.2 Entire Proteome of *S. aureus*

As genomics serves as the backbone of proteomics, publication of complete genome sequence of *S. aureus* in the year 2001 laid the foundation for *S. aureus* proteomics [65]. From the genome sequence, the number of open reading frames was predicted to be around 2600. Comprehensive mass spectrometric studies coupled with two-dimensional polyacrylamide gel electrophoresis (2-DGE) identified 1123 cytoplasmic proteins which represent 66% of predicated cytoplasmic proteins, and 2-DGE reference map (with 473 identified proteins) of *S. aureus* cellular proteome was first established in 2005 [66]. Later, the total proteome of *S. aureus* comprising cytoplasmic, surface-associated, membrane, and secreted proteome was predicted to be 2618 proteins of which 2005 proteins (77%) have been identified (Fig. 18.2) [67].

18.5.3 Proteomic Strategies of *S. aureus*

Proteomic strategies are basically composed of protein extraction, purification, separation, and identification. Gel-based separations of proteins are common and widely used in the field of comparative proteomics. Advancements in the mass

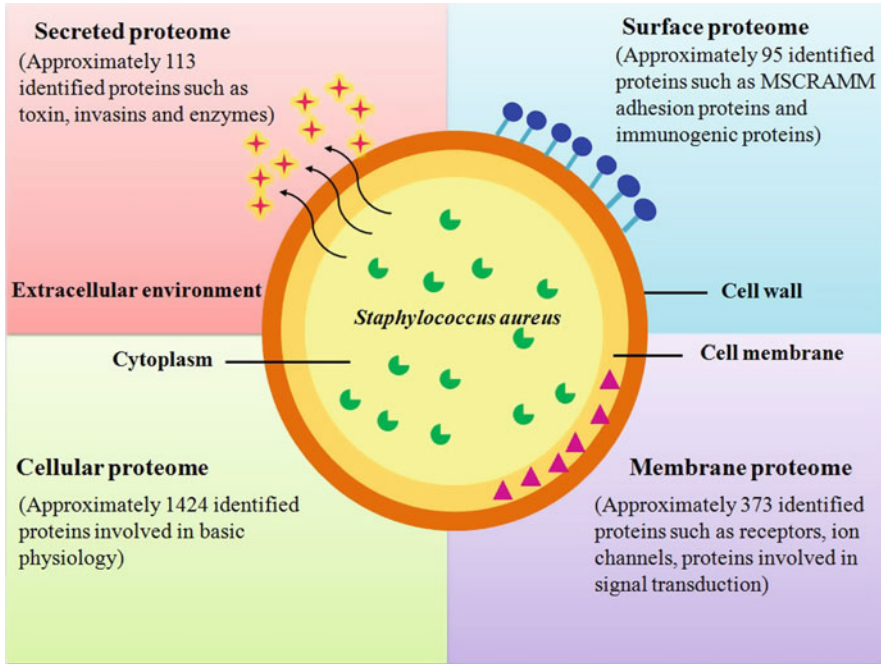


Fig. 18.2 Compendious proteome of *S. aureus* (Data obtained from [67])

spectrometric technologies enabled the gel-free quantification of proteins based on spectral counting and peak intensities [68]. Comprehensive workflow of *S. aureus* proteomic strategies is presented in Fig. 18.3.

One-dimensional SDS-polyacrylamide gel electrophoresis (SDS-PAGE) is the simplest gel-based proteomic technique used for the separation of proteins according to molecular weight. In case of crude protein samples, this technique is used for the purification of proteins prior to further analysis. Native PAGE analysis is generally used to identify the known protein targets in native form [69]. In 1975, 2-DGE was first introduced by O'Farrell and Klose and remains a gold standard proteomic technique for the separation of complex protein mixtures till date [70]. The workflow of 2-DGE comprises extraction and purification of proteins, rehydration, first dimensional separation based on isoelectric point otherwise known as isoelectric focusing, reduction, alkylation, second dimensional separation based on molecular weight, staining and visualization of protein spots, image analysis and in-gel digestion of proteins, mass spectrometry, and database search based identification [71]. Advancement of 2-DGE with the use of mass spectrometry compatible CyDyes led to an effective proteomic approach difference gel electrophoresis (DIGE). This technique excludes the gel-to-gel variations which are main disadvantage of 2-DGE and also provides extensive relative quantification of proteins [72]. Comparative gel-based analysis of protein samples from control and treated cells can identify the

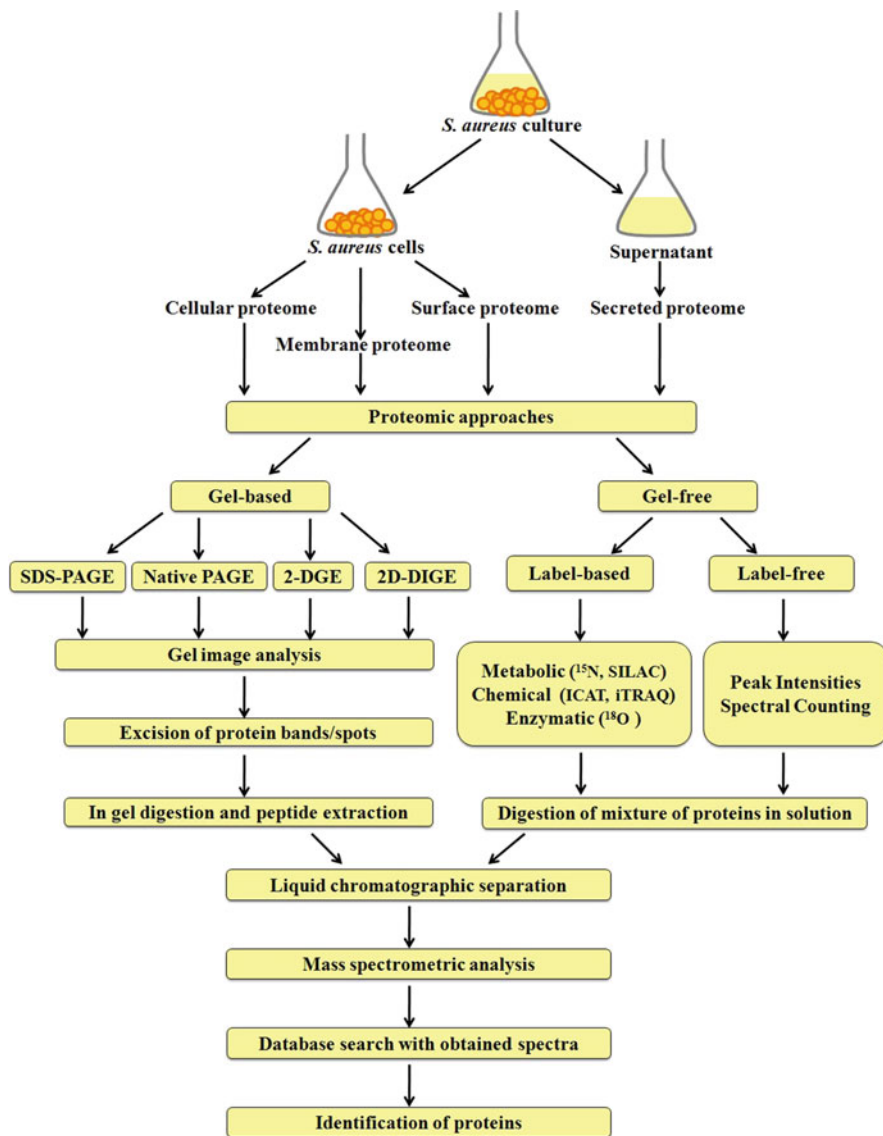


Fig. 18.3 Schematic illustration of comprehensive workflow involved in various strategies of *S. aureus* proteomics

differentially regulated proteins, and mass spectrometric identification of proteins spots can reveal the molecular protein targets of drugs [73].

2-DGE-based proteomic study revealed that rhodomyrtone interrupted cell wall biosynthesis and cell division in *S. aureus* to exert antibacterial activity [74]. Our previous study identified the multiple protein targets of citral to inhibit biofilm and

virulence. Citral upregulated the transcriptional repressor CodY which suppresses the major adhesion and secreted virulence factors [75]. 2-DGE-based proteomic analysis of cellular proteins of *S. aureus* treated with juglone exhibited the inhibition of DNA and RNA synthesis [76]. Quantitative proteomic analysis using isobaric tags unveiled inhibition of protein synthesis by 3-O-alpha-L-(2'',3''-di-p-coumaroyl) rhamnoside in *S. aureus* [77]. Spectral counting-based label-free quantitative proteomics of oxacillin-treated *S. aureus* revealed the upregulation of tolerance and resistance mechanisms [78]. Disruption of oxidation-reduction homeostasis and cell wall biosynthesis by combination of erythromycin and oxacillin was elucidated by spectral counting-based label-free proteomic approach [79]. Disruption of iron homeostasis induces SOS response in *S. aureus* upon treatment with punicalagin identified from pomegranate through quantitative isobaric labeling-based proteomic approach [80].

18.6 Concluding Remarks

This chapter demonstrated the pathogenesis, major virulence determinants, and biofilm formation of *S. aureus* and its clinical relevance. Alternative therapeutic developments of drug discovery to overcome the burden of antibiotic resistance are provided in detail. The importance of proteomics in the field of drug discovery and target identification and various proteomic strategies including gel-based and gel-free techniques in aspect of decoding the molecular targets of drugs are discussed. Understanding of *S. aureus* pathogenesis and current approaches for drug target identification will serve as platform for future studies for the development of effective strategies to combat *S. aureus* infections.

Acknowledgment The authors sincerely acknowledge DST-FIST [Grant No. SR/FST/LSI-639/2015(C)], UGC-SAP [Grant No. F.5-1/2018/DRS-II(SAP-II)] and DST-PURSE [Grant No. SR/PURSE Phase 2/38 (G)] for providing instrumentation facilities. The authors also thank RUSA 2.0 [F.24-51/2014-U, Policy (TN Multi-Gen), Dept. of Edn, GoI], and SKP is thankful to UGC for Mid-Career Award [F.19-225/2018(BSR)].

Competing Interest

All the authors declare no conflict of interest.

References

1. Fitzpatrick F, Humphreys H, O'gara JP (2005) The genetics of staphylococcal biofilm formation—will a greater understanding of pathogenesis lead to better management of device-related infection? *Clin Microbiol Infect* 11(12):967–973
2. Bhattacharya M, Wozniak DJ, Stoodley P, Hall-Stoodley L (2015) Prevention and treatment of *Staphylococcus aureus* biofilms. *Expert Rev Anti-Infect Ther* 13(12):1499–1516
3. Bonar E, Wójcik I, Władyska B (2015) Proteomics in studies of *Staphylococcus aureus* virulence. *Acta Biochim Pol* 62(3):367–381

4. Wertheim HF, Melles DC, Vos MC, van Leeuwen W, van Belkum A, Verbrugh HA, Nouwen JL (2005) The role of nasal carriage in *Staphylococcus aureus* infections. *Lancet Infect Dis* 5(12):751–762
5. Sakr A, Brégeon F, Mège JL, Rolain JM, Blin O (2018) *Staphylococcus aureus* nasal colonization: an update on mechanisms, epidemiology, risk factors, and subsequent infections. *Front Microbiol* 9:2419
6. Kobayashi SD, Malachowa N, DeLeo FR (2015) Pathogenesis of *Staphylococcus aureus* abscesses. *Am J Pathol* 185(6):1518–1527
7. Grundmann H, Aanensen DM, Van Den Wijngaard CC, Spratt BG, Harmsen D, Friedrich AW, European Staphylococcal Reference Laboratory Working Group (2010) Geographic distribution of *Staphylococcus aureus* causing invasive infections in Europe: a molecular-epidemiological analysis. *PLoS Med* 7(1):e1000215
8. Klevens RM, Morrison MA, Nadle J, Petit S, Gershman K, Ray S, Harrison LH, Lynfield R, Dumyati G, Townes JM, Craig AS (2007) Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States. *JAMA* 298(15):1763–1771
9. Otter JA, French GL (2011) Community-associated methicillin-resistant *Staphylococcus aureus* strains as a cause of healthcare-associated infection. *J Hosp Infect* 79(3):189–193
10. Pinto RM, Lopes-de-Campos D, Martins MCL, Van Dijk P, Nunes C, Reis S (2019) Impact of nanosystems in *Staphylococcus aureus* biofilms treatment. *FEMS Microbiol Rev* 43(6):622–641
11. Schierholz JM, Beuth J (2001) Implant infections: a haven for opportunistic bacteria. *J Hosp Infect* 49(2):87–93
12. Tong SY, Davis JS, Eichenberger E, Holland TL, Fowler VG (2015) *Staphylococcus aureus* infections: epidemiology, pathophysiology, clinical manifestations, and management. *Clin Microbiol Rev* 28(3):603–661
13. McGavin MJ, Heinrichs DE (2012) The staphylococci and staphylococcal pathogenesis. *Front Cell Infect Microbiol* 2:66
14. Tacconelli E, Magrini N, Kahlmeter G, Singh N (2017) Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. *World Health Organ* 27:318–327
15. Centres for Disease Control and Prevention (US) (2013) Antibiotic resistance threats in the United States, 2013. Centres for Disease Control and Prevention, US Department of Health and Human Services
16. Bennett JW, Chung KT (2001) Alexander Fleming and the discovery of penicillin
17. Fleming A (1929) On the antibacterial action of cultures of a *Penicillium*, with special reference to their use in the isolation of *B. influenzae*. *Br J Exp Pathol* 10(3):226
18. Gaynes R (2017) The discovery of penicillin—new insights after more than 75 years of clinical use. *Emerg Infect Dis* 23(5):849
19. Khan MF (2017) Brief history of *Staphylococcus aureus*: a focus to antibiotic resistance. *EC Microbiol* 5(2):36–39
20. Abraham EP, Chain E (1940) An enzyme from bacteria able to destroy penicillin. *Nature* 146(3713):837–837
21. Fleming A (1945) Penicillin. Nobel Lecture, December 11, 1945. Nobel e-museum
22. Kim J (2009) Understanding the evolution of methicillin-resistant *Staphylococcus aureus*. *Clin Microbiol News* 31(3):17–23
23. Hiramatsu K, Hanaki H, Ino T, Yabuta K, Oguri T, Tenover FC (1997) Methicillin-resistant *Staphylococcus aureus* clinical strain with reduced vancomycin susceptibility. *J Antimicrob Chemother* 40(1):135–136
24. Centers for Disease Control and Prevention (CDC) (2002) *Staphylococcus aureus* resistant to vancomycin – United States, 2002. *MMWR. Morbidity and mortality weekly report*, 51(26), p. 565

25. Tsioupras S, Gold HS, Sakoulas G, Eliopoulos GM, Wennersten C, Venkataraman L, Moellering RC Jr, Ferraro MJ (2001) Linezolid resistance in a clinical isolate of *Staphylococcus aureus*. *Lancet* 358(9277):207–208
26. Gould IM (2006) Costs of hospital-acquired methicillin-resistant *Staphylococcus aureus* (MRSA) and its control. *Int J Antimicrob Agents* 28(5):379–384
27. Zucca M, Savoia D (2010) The post-antibiotic era: promising developments in the therapy of infectious diseases. *Int J Biomed Sci* 6(2):77
28. Motallebi M, Alibolandi Z, Aghmiyuni ZF, van Leeuwen WB, Sharif MR, Moniri R (2020) Molecular analysis and the toxin, MSCRAMM, and biofilm genes of methicillin-resistant *Staphylococcus aureus* strains isolated from pemphigus wounds: a study based on SCCmec and dru typing. *Infect Genet Evol* 87:104644
29. Gross M, Cramton SE, Götz F, Peschel A (2001) Key role of teichoic acid net charge in *Staphylococcus aureus* colonization of artificial surfaces. *Infect Immun* 69(5):3423–3426
30. Cue DR, Lei MG, Lee C (2012) Genetic regulation of the intercellular adhesion locus in staphylococci. *Front Cell Infect Microbiol* 2:38
31. Hong X, Qin J, Li T, Dai Y, Wang Y, Liu Q, He L, Lu H, Gao Q, Lin Y, Li M (2016) Staphylococcal protein A promotes colonization and immune evasion of the epidemic healthcare-associated MRSA ST239. *Front Microbiol* 7:951
32. Mirzaee M, Najjar-Peeraeyeh S, Behmanesh M (2015) Prevalence of fibronectin-binding protein (FnA and FnB) genes among clinical isolates of methicillin resistant *Staphylococcus aureus*. *Mol Genet Microbiol Virol* 30(4):221–224
33. Park PW, Roberts DD, Grosso LE, Parks WC, Rosenbloom J, Abrams WR, Mechem RP (1991) Binding of elastin to *Staphylococcus aureus*. *J Biol Chem* 266(34):23399–23406
34. Hudson MC, Ramp WK, Frankenburg KP (1999) *Staphylococcus aureus* adhesion to bone matrix and bone-associated biomaterials. *FEMS Microbiol Lett* 173(2):279–284
35. Herman-Bausier P, Labate C, Towell AM, Derclaye S, Geoghegan JA, Dufrêne YF (2018) *Staphylococcus aureus* clumping factor A is a force-sensitive molecular switch that activates bacterial adhesion. *Proc Natl Acad Sci* 115(21):5564–5569
36. Porayath C, Suresh MK, Biswas R, Nair BG, Mishra N, Pal S (2018) Autolysin mediated adherence of *Staphylococcus aureus* with Fibronectin, Gelatin and Heparin. *Int J Biol Macromol* 110:179–184
37. Laarman AJ, Ruyken M, Malone CL, van Strijp JA, Horswill AR, Rooijackers SH (2011) *Staphylococcus aureus* metalloprotease aureolysin cleaves complement C3 to mediate immune evasion. *J Immunol* 186(11):6445–6453
38. Stach N, Kaszycki P, Władyka B, Dubin G (2018) Extracellular proteases of *Staphylococcus* spp. In: Pet-to-man travelling staphylococci. Academic Press, London, pp 135–145
39. Hu C, Xiong N, Zhang Y, Rayner S, Chen S (2012) Functional characterization of lipase in the pathogenesis of *Staphylococcus aureus*. *Biochem Biophys Res Commun* 419(4):617–620
40. Kiedrowski MR, Crosby HA, Hernandez FJ, Malone CL, McNamara JO II, Horswill AR (2014) *Staphylococcus aureus* Nuc2 is a functional, surface-attached extracellular nuclease. *PLoS One* 9(4):e95574
41. Mandell GL (1975) Catalase, superoxide dismutase, and virulence of *Staphylococcus aureus*. In vitro and in vivo studies with emphasis on staphylococcal-leukocyte interaction. *J Clin Invest* 55(3):561–566
42. Ibberson CB, Jones CL, Singh S, Wise MC, Hart ME, Zurawski DV, Horswill AR (2014) *Staphylococcus aureus* hyaluronidase is a CodY-regulated virulence factor. *Infect Immun* 82(10):4253–4264
43. Cheng AG, McAdow M, Kim HK, Bae T, Missiakas DM, Schneewind O (2010) Contribution of coagulases towards *Staphylococcus aureus* disease and protective immunity. *PLoS Pathog* 6(8):e1001036
44. Periasamy S, Chatterjee SS, Cheung GY, Otto M (2012) Phenol-soluble modulins in staphylococci: what are they originally for? *Commun Integr Biol* 5(3):275–277

45. Vandenesch F, Lina G, Henry T (2012) *Staphylococcus aureus* hemolysins, bi-component leukocidins, and cytolytic peptides: a redundant arsenal of membrane-damaging virulence factors? *Front Cell Infect Microbiol* 2:12
46. Argudín MÁ, Mendoza MC, Rodicio MR (2010) Food poisoning and *Staphylococcus aureus* enterotoxins. *Toxins* 2(7):1751–1773
47. Shallcross LJ, Fragaszy E, Johnson AM, Hayward AC (2013) The role of the Pantone-Valentine leukocidin toxin in staphylococcal disease: a systematic review and meta-analysis. *Lancet Infect Dis* 13(1):43–54
48. Kulhankova K, Kinney KJ, Stach JM, Gourronc FA, Grumbach IM, Klingelutz AJ, Salgado-Pabón W (2018) The superantigen toxic shock syndrome toxin 1 alters human aortic endothelial cell function. *Infect Immun* 86(3):e00848-17
49. Bien J, Sokolova O, Bozko P (2011) Characterization of virulence factors of *Staphylococcus aureus*: novel function of known virulence factors that are implicated in activation of airway epithelial proinflammatory response. *J Pathog* 2011: 601905, pp. 1–13
50. Oogai Y, Matsuo M, Hashimoto M, Kato F, Sugai M, Komatsuzawa H (2011) Expression of virulence factors by *Staphylococcus aureus* grown in serum. *Appl Environ Microbiol* 77(22):8097–8105
51. Gordon RJ, Lowy FD (2008) Pathogenesis of methicillin-resistant *Staphylococcus aureus* infection. *Clin Infect Dis* 46(Supplement_5):S350–S359
52. Yarwood JM, Bartels DJ, Volper EM, Greenberg EP (2004) Quorum sensing in *Staphylococcus aureus* biofilms. *J Bacteriol* 186(6):1838–1850
53. Kierek-Pearson K, Karatan E (2005) Biofilm development in bacteria. *Adv Appl Microbiol* 57:79–111
54. Arciola CR, Campoccia D, Montanaro L (2018) Implant infections: adhesion, biofilm formation and immune evasion. *Nat Rev Microbiol* 16(7):397
55. Archer NK, Mazaitis MJ, Costerton JW, Leid JG, Powers ME, Shirtliff ME (2011) *Staphylococcus aureus* biofilms: properties, regulation, and roles in human disease. *Virulence* 2(5):445–459
56. Høiby N, Bjarnsholt T, Givskov M, Molin S, Ciofu O (2010) Antibiotic resistance of bacterial biofilms. *Int J Antimicrob Agents* 35(4):322–332
57. Arciola CR, Campoccia D, Speziale P, Montanaro L, Costerton JW (2012) Biofilm formation in *Staphylococcus* implant infections. A review of molecular mechanisms and implications for biofilm-resistant materials. *Biomaterials* 33(26):5967–5982
58. Lister JL, Horswill AR (2014) *Staphylococcus aureus* biofilms: recent developments in biofilm dispersal. *Front Cell Infect Microbiol* 4:178
59. Waters CM, Bassler BL (2005) Quorum sensing: cell-to-cell communication in bacteria. *Annu Rev Cell Dev Biol* 21:319–346
60. Cerca N, Brooks JL, Jefferson KK (2008) Regulation of the intercellular adhesin locus regulator (icaR) by SarA, σ B, and IcaR in *Staphylococcus aureus*. *J Bacteriol* 190(19):6530–6533
61. François P, Scherl A, Hochstrasser D, Schrenzel J (2010) Proteomic approaches to study *Staphylococcus aureus* pathogenesis. *J Proteome* 73(4):701–708
62. Sleno L, Emili A (2008) Proteomic methods for drug target discovery. *Curr Opin Chem Biol* 12(1):46–54
63. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature* 422:198–207
64. Savaryn JP, Catherman AD, Thomas PM, Abecassis MM, Kelleher NL (2013) The emergence of top-down proteomics in clinical research. *Genome Med* 5(6):1–8
65. Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki KI, Nagai Y, Lian J (2001) Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*. *Lancet* 357(9264):1225–1240
66. Kohler C, Wolff S, Albrecht D, Fuchs S, Becher D, Büttner K, Engelmann S, Hecker M (2005) Proteome analyses of *Staphylococcus aureus* in growing and non-growing cells: a physiological approach. *Int J Med Microbiol* 295(8):547–565

67. Hecker M, Mäder U, Völker U (2018) From the genome sequence via the proteome to cell physiology—Pathoproteomics and pathophysiology of *Staphylococcus aureus*. *Int J Med Microbiol* 308(6):545–557
68. Chandramouli K, Qian PY (2009) Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum Genomics Proteomics* 2009:239204
69. Nowakowski AB, Wobig WJ, Petering DH (2014) Native SDS-PAGE: high resolution electrophoretic separation of proteins with retention of native properties including bound metal ions. *Metallomics* 6(5):1068–1078
70. Oliveira BM, Coorssen JR, Martins-de-Souza D (2014) 2DE: the phoenix of proteomics. *J Proteome* 104:140–150
71. Buyukkoroglu G, Dora DD, Özdemir F, Hizel C (2018) Techniques for protein analysis. In: *Omics technologies and bio-engineering*. Academic Press, London, pp 317–351
72. Magdeldin S, Enany S, Yoshida Y, Xu B, Zhang Y, Zureena Z, Lokamani I, Yaoita E, Yamamoto T (2014) Basics and recent advances of two dimensional-polyacrylamide gel electrophoresis. *Clin Proteomics* 11(1):1–10
73. Penque D (2009) Two-dimensional gel electrophoresis and mass spectrometry for biomarker discovery. *Proteomics Clin Appl* 3(2):155–172
74. Sianglum W, Srimanote P, Wonglumsom W, Kittiniyom K, Voravuthikunchai SP (2011) Proteome analyses of cellular proteins in methicillin-resistant *Staphylococcus aureus* treated with rhodomymrtone, a novel antibiotic candidate. *PLoS One* 6(2):e16628
75. Valliammai A, Sethupathy S, Ananthi S, Priya A, Selvaraj A, Nivetha V, Aravindraja C, Mahalingam S, Pandian SK (2020) Proteomic profiling unveils citral modulating expression of IsaA, CodY and SaeS to inhibit biofilm and virulence in methicillin-resistant *Staphylococcus aureus*. *Int J Biol Macromol* 158:208–221
76. Wang J, Wang Z, Wu R, Jiang D, Bai B, Tan D, Yan T, Sun X, Zhang Q, Wu Z (2016) Proteomic analysis of the antibacterial mechanism of action of Juglone against *Staphylococcus aureus*. *Nat Prod Commun* 11(6):1934578X1601100632
77. Carruthers NJ, Stemmer PM, Media J, Swartz K, Wang X, Aube N, Hamann MT, Valeriote F, Shaw J (2020) The anti-MRSA compound 3-O-alpha-L-(2'',3''-di-p-coumaroyl) rhamnoside (KCR) inhibits protein synthesis in *Staphylococcus aureus*. *J Proteome* 210:103539
78. Liu X, Hu Y, Pai PJ, Chen D, Lam H (2014) Label-free quantitative proteomics analysis of antibiotic response in *Staphylococcus aureus* to oxacillin. *J Proteome Res* 13(3):1223–1233
79. Liu X, Pai PJ, Zhang W, Hu Y, Dong X, Qian PY, Chen D, Lam H (2016) Proteomic response of methicillin-resistant *S. aureus* to a synergistic antibacterial drug combination: a novel erythromycin derivative and oxacillin. *Sci Rep* 6:19841
80. Cooper B, Islam N, Xu Y, Beard HS, Garrett WM, Gu G, Nou X (2018) Quantitative proteomic analysis of *Staphylococcus aureus* treated with punicalagin, a natural antibiotic from pomegranate that disrupts iron homeostasis and induces SOS. *Proteomics* 18(9):1700461



Proteomics in the Study of Host-Pathogen Interactions

19

Preethi Sudhakara, S. Kumaran, and Wilson Aruni

Abstract

DNA is referred to as the basic blueprint of life, but the implementation of the genetic plan is carried out by the activities of proteins. Hence, the biological diversity that is noticed in nature is therefore protein-based, and protein-level modifications pave way for the natural selection. Proteomics is the study of the proteome. The proteome is the genome-operating process that delineates the array of proteins that are produced in biological compartments such as cell, tissue or organs at a specific time, under specific conditions. Proteomics act as a bridge between our understanding of genomic study and cellular processes. Proteomics proposes a conspicuous method to study the proteomes relationship between the host and pathogen during their complex biochemical cross talk. From the comparably small number of genes in every organism whether the host or the pathogen, the proteome stands big owing to many characteristics such as the slicing event, single gene-multiprotein process, posttranslational modifications, etc. These processes that follow make proteomics an indispensable omics to study the host and pathogen interactions in any system. Pathogenic diseases are a result of host-pathogen interaction which encompasses molecular “cross talk.” This chapter reviews a gist of major important aspects of proteomics in determining the host-pathogen interactome. This overview aims to understand the various proteomic

P. Sudhakara

College of Medicine, University of Florida, Gainesville, USA

S. Kumaran

Sathyabama Institute of Science and Technology (Deemed University), Chennai, India

W. Aruni (✉)

Sathyabama Institute of Science and Technology (Deemed University), Chennai, India

US Department of Veteran Affairs, Loma Linda VA, Loma Linda, USA

e-mail: Aruni.wilson@rccd.edu; drwilsonprovcc@sathyabama.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_19

341

techniques, thereby paving way for their application to study the molecular mechanisms of bioactive molecules, characterize complex protein networks, find (bio)marker proteins for diagnosis of diseases, and characterize targets for formulating the drugs for pharmaceuticals.

Keywords

Proteomics · Host pathogen interaction · Interactomics · Metabolomics

19.1 Introduction

Pathogens remain a potential threat to the animal kingdom especially to the human health and well-being. Pathogens subvert the host immune mechanisms and hamper various cellular processes. Relationship between the host and pathogen represents equilibrium between mechanisms of virulence modulation and host defense processes [1].

Despite the encoded message of life process is through DNA, the encoding proteins carryout the functions. Hence, the biological diversity is therefore protein-based [2]. The proteome is pivotal in reacting and stabilizing the cells to environmental signals [3]. It comprises a first step of the cascade, the cyto-sensorium (such as the cellular process proteins that are sensors, receptors, and message transfer units from environmental signals), and the following sequence, the cyto-effectorium (i.e., either individual proteins or a group of proteins in response to environmental changes).

Proteomics is the study of the proteome, meaning all the proteins produced by a cell or tissue. Proteomics paves the way in understanding of genome sequence and their specific cellular behavior. This offers a tool to study the reaction of the host and pathogen proteomes (i.e., genome-operating systems) during their complex biochemical cross talk [4].

Advances in proteomic technologies afford opportunities to compare protein content between biological systems. Hence, this technique is useful to explore and characterize the host-pathogen interactions from a global proteomic view. The technique is mainly used for two important processes, namely, determination of the content of protein and also cataloguing the proteomic alteration due to host interaction with the microbe in question. The recent proteomics methods are mainly used to obtain in-depth knowledge on the mechanistic role of pathogen virulence and hence play a pivotal role in identifying pathogen-specific markers [5].

The adaptation of omics technologies and their in-depth analysis of cells, tissues, and organisms gives a comprehended picture of their role in the system. They mainly include the detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics). This multifaceted study is referred to as high-dimensional biology. While many areas of omics can be integrated, focusing on a proteomics-based interaction study would be advantageous as the system will be mined for the end product of all the omics, that is, “proteins.” From the comparably

small number of genes in every organisms whether the host or the pathogen, the proteome stands big owing to many characteristics such as the slicing event, single gene-multiprotein process, posttranslational modifications, etc. These processes that follow make proteomics an indispensable omics to study the host and pathogen interactions in any system.

Systems biology and omics differ from traditional studies, which are largely hypothesis-driven [6]. They have many applications. “Omics” technology can be applied for understanding of normal physiological processes, disease progression, and diagnosis and prognosis of a disease. Proteomics characterize the relevant information within the cell and the organism, through protein pathways and protein and metabolic networks [7, 8], hence facilitating to understand the functional relevance of proteins [9]. However, the drawback of studying proteomics is the complicated domain size (>100,000 proteins) and its inability to detect accurately low-abundance proteins. The proteome is a dynamic reflection of both genes and the environment. However, the practical bottlenecks are protein concentration, sample purification, and digestion procedures.

19.2 Host-Pathogen Interaction

The intimate relationship between host and pathogen leads to a diversified molecular cross talk resulting in the process of any disease. This complex molecular interaction and dialogue between host and pathogen are much needed in order to explore our understanding of pathogen virulence and also to develop pathogen-specific proteome. Most host species have acquired strategies by selective pressure to mislead the pathogen and gain the host system through molecular dialogue between the two components. However, many pathogen species have acquired various strategies through selective pressure in order to bypass the host defenses and winning the molecular war and to complete its life cycle. Always, pathogens remain a significant threat to any host species. Hence it is critical that the ability to detect, treat, and contain transmission is an important sequential process.

Proteomics applications in studying the host-pathogen interactions are in their developing stages, and the recent techniques could lead to new insights on host specificity, pathogen evolution, and pathogen virulence. Many new conceptual approaches to decipher host-pathogen interactions open up new avenues to determine the cross talk diversity involved in trophic interactions during host-pathogen interactions.

Host-pathogen interactions are indispensable for the study of disease progression. Hence using proteomics tools to analyze different stages of infection, pathogen invasion and proliferation in their hosts could enlighten complex network processes and intricate pathways of virulence determination (Fig. 19.1). Proteomics approach coupled with bioinformatics continues to play a large role in expanding our knowledge in the field of proteomics, through datasets, gene and genome, protein alignment methods, structural analysis, and machine learning methods hence vital for the

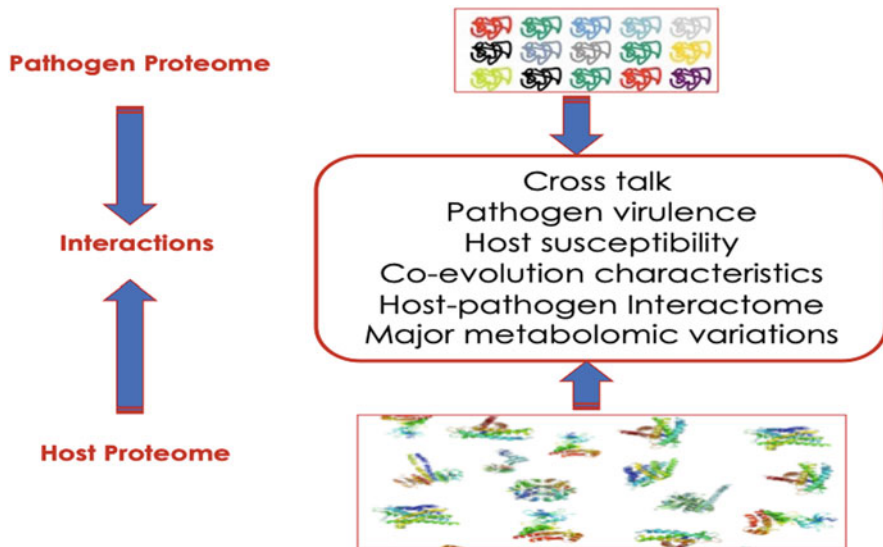


Fig. 19.1 Major interactions of pathogen and host proteome

understanding of host-pathogen interactions and their implication in the treatment regimen.

19.3 Methods in Proteomic Study

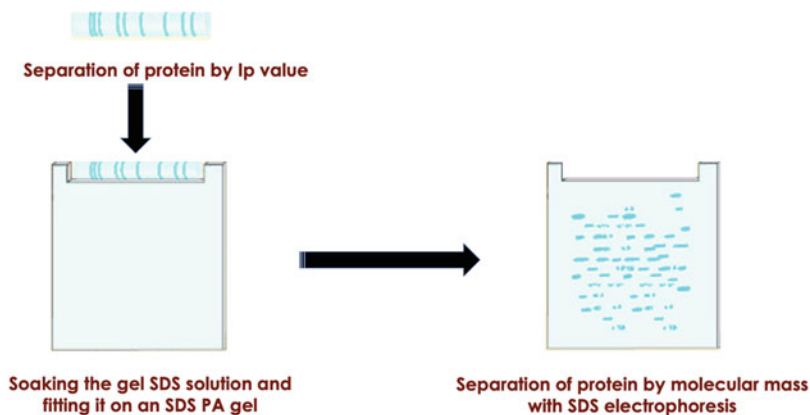
Among the various methods, the first-generation proteomics approach, a one-dimensional gel electrophoresis, was initially used to study the overall major protein expression between the host and pathogen interaction. However, with its inherent drawbacks, the other two major methods, namely, two-dimensional electrophoresis (2-DE) and mass spectrometry (MS), can explore the posttranslational modifications of host and pathogen proteins (such as phosphorylation, glycosylation, acetylation, and methylation). Such mechanisms are considered as major cellular regulation in health and diseases. Although 2-DE offers a high-quality approach for the study of host and pathogen proteomes, several proteomics tools help support this approach for data analysis [10]. Table 18.1 shows a comparison of the most popular proteomics tools.

19.4 In-Gel Proteomics

One of the most common and primitive methods of proteomic study include the one-dimensional gel electrophoresis. This process includes separation of proteins based on their charge in an electric gradient. However, the drawbacks include minute

Table 18.1 A comparison of proteomics tools

Name of the technique	Separation
2-DE	Electrophoresis: IEF PAGE
2-DIGE	Electrophoresis: IEF PAGE
MuDPIT	LC-LC of peptides
ICAT	LC of peptides
SELDI-TOF-MS	Physio-chemical characteristics
Protein arrays	Antibody-based affinity binding reagents

**Fig. 19.2** 2D gel electrophoresis—principle. Separation of protein based on their p values and molecular mass using SDS-PAGE

mobility variations which cannot be much appreciated in such a study. However, a two-dimensional electrophoresis (Fig. 19.2) method is used to separate proteins employing two different properties of the protein [11].

One of the key tools for comparative proteomics research is the two-dimensional gel electrophoresis (2-DE). In 2-DE, mixtures of proteins are separated by charge (isoelectric point, pI) in the first dimension and separated by mass in the second dimension.

The applicability and adoptability of this method were enhanced because of the introduction of immobilized pH gradient strips, because this technique showed good reproducible results and handling became easy. The 2-DE can be achieved through the separation of several thousands of different proteins in one gel. Stain techniques, namely, Coomassie Brilliant Blue, silver, SYPRO Ruby, and Deep Purple, can be employed to visualize the proteins. However, this method can detect proteins in the size range of 10–200 kDa at pH 3.5–11.5 and also, ineffective at distinguishing between low-abundant proteins and small molecular weight proteins (<10 kDa). Various quantification methods and fluorescent dyes with multiplexing approach are now being replaced by mass spectrometric methods. This technique is more used to

study variations between two systems or pathogens or their strains and their interaction with the same host cell [12].

19.5 DIGE

The ultimate evolution of 2-DE technique, DIGE, significantly improves the analytical power of gel-based methods in proteome research. Proteome samples that are previously labeled with spectrally resolvable fluorophore agents (CyDyes™: Cy2, Cy3, and Cy5) lead to a more accurate normalization of protein spots. Each specific fluorophore is excited, generating unique gel images corresponding to each pre-labeled proteome sample. DIGE can also be used together with gel-free methods (e.g., LC/MS, LC-MS/MS), which will improve the analytical power.

In order to make this technique more specific, protein degradation (e.g., hydrolysis and oxidation) and modification (e.g., carbamylation) must be minimized. Membrane proteins are difficult to resolve due to **hydrophobicity** and **lipid bilayers**. Table 18.2 shows the common protein staining methods.

The major methods followed in studying the host-pathogen interaction proteomics are in-gel proteomics and gel-free proteomics methods. However, the two major analytical processes in the proteomic study follows a sequence of sample preparation, an analytical platform, separation, detection, and identification through bioinformatics validation using databases. The two major analytical processes in the study of proteomics are given in Fig. 19.3.

19.6 Host-Pathogen Interaction Study Using Various Proteomics Methods

Proteomics plays a crucial role, allowing the discovery of disease biology and mechanisms to identify new drug targets and much more. Applications for proteomics in the drug discovery process, for instance, include the detection of human and animal disease biomarkers for which studying the global proteome variation between the infected and healthy host is indispensable. Hence, in order to study the

Table 18.2 Common protein stains used in 2D gel electrophoresis

Staining method	Dye
Post-electrophoretic stains	
Coomassie	Coomassie Brilliant Blue-R250
Bright blue	Coomassie Brilliant Blue-G250
Silver stain	Silver nitrate or silver ammonia
Negative	Zinc and or imidazole
Fluorescent	SYPRO Ruby, SYPRO Orange, Red and Tangerine, Epicoccone
Pre-electrophoretic stains	
Fluorescence	CyDyes, FlaSHPro Dyes

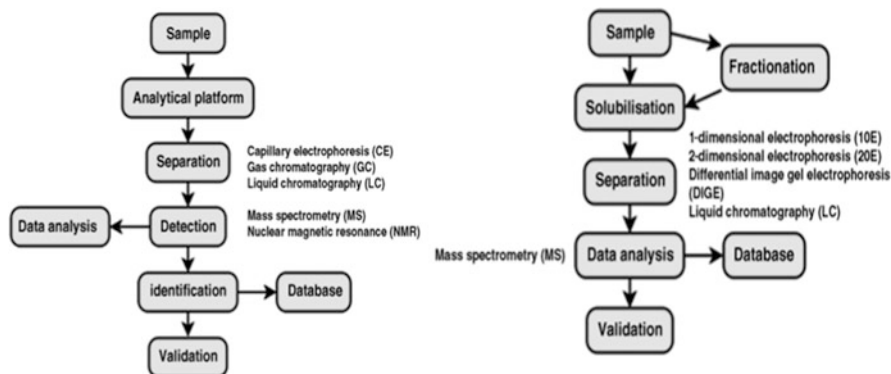


Fig. 19.3 Two major processes of proteomic study given as a sequential flow diagram

pharmacological responses to therapeutic treatments, identification of specific biomarkers can be widely described as measures of normal biological or pathologic disease states. Currently, more predictive tools need to be developed to assist in the early evaluation of host-pathogen interaction cycle at various stages of infection, as well as to better understand disease models and their complex interwoven pathways between the host and pathogen. Currently, biomarkers for many human diseases are insufficient for early detection of the disease and lack sensitivity or specificity. Identifying earlier and more accurate disease biomarkers will significantly increase the options for medical care and the likelihood of success for which proteomic study plays a major role.

To identify both potential disease mechanisms and disease biomarkers, proteomic-based platforms are becoming increasingly strong with precision. For biological understanding at a wide-scale stage, proteomics involves using highly complex protein screening technology. This knowledge can then be used to provide an interpretation of the fundamental biological processes underlying diseases in conjunction with other “omics” data. In the last decade, the advent of newer sophisticated mass spectrometry (MS) technology, with higher resolution and faster scan speeds, has made it possible to classify highly complex proteomes easier and faster with shorter analysis periods [13].

One of the most important and pivotal studied processes is the host-pathogen interaction and has focused on newer and yet uncultivable pathogens of human microbiome and is currently being studied for their proteome variations during interaction (interactomes). Nevertheless, simultaneous host proteome modulations are studied through in-gel and/or direct proteomic studies using MS analysis.

Among the many microbiome keystone pathogen players that shift the entire niche from symbiosis to dysbiosis, one of the major keystone pathogens of human oral microbiome *Filifactor alocis* and its host interactions were studied. A comparison of one dimension, two dimension, and MS analysis is given hereunder.

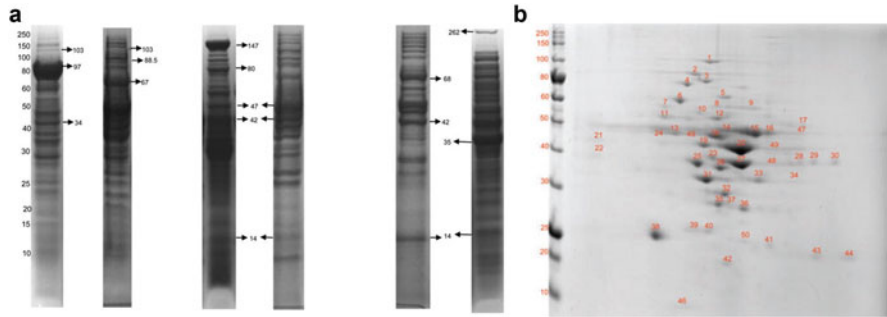


Fig. 19.4 (a) One-dimensional gel electrophoresis of five different *F. alocis* strains of pathogen proteomes. (b) Two-dimensional gel electrophoresis of *F. alocis* strain after interaction with the host

Figure 19.4a shows the one-dimensional protein mobility of various strains of *Filifactor alocis* showing an arbitrary variation among the strains and variations among the cellular and cell-free fractions of the bacteria. An in-depth study using 2D gel electrophoresis showed a wide variation among the strains also before and after interaction with the host (Fig. 19.4b) [14].

An SDS-PAGE analysis of cell fractions from *F. alocis* ATCC 35896 strain and the D-62D strains on comparison in order to study the strain variations in virulence and interaction showed variation in their protein profile (Fig. 19.4b). Further to MS/MS analysis of proteins spots, a total of approximately 1568 peptides were identified that were above the threshold ($p < 0.05$) and had a Mascot score of ≥ 15 with individual ion score of more than 20. A total of 986 nonredundant peptides corresponding to 219 proteins were identified for each protein. Thus in-gel proteomics approach was used to study the major host-pathogen modulatory pathways [14].

In continuation, in-gel proteomic study of the same samples was performed. The expressed 2D spots from the gel were excised and were trypsin digested [14] and were subjected to mass spectrometry analysis. The study could determine unique expressive proteins during their interaction. The proteomic results through mass spectrometry identified novel and unique proteomes that are mainly involved in amino acid metabolism, secretory system, and virulence determination. The relative abundance of such protein on comparison gave a conclusive picture on the unique virulence determining pathways and their sequence of protein interactions. One of the unique findings of this technique was the identification of important proteins that have moonlighting function in virulence. Hence, the in-gel proteomics approach deciphered on many novel such proteins unique to the bacterium as well as other general proteins found to possess similar function in other human pathogens.

19.7 Gel-Based Autoradiography and Chemiluminescence

19.7.1 Autoradiography

Labeling of proteins either prior to or post-electrophoretically using radioactive isotopes remains the most sensitive method for protein detection. Individual radiolabeled protein bands or spots are usually detected in one of three ways: liquid scintillation counting, autoradiography, and fluorography. Modifications to facilitate the detection of proteins expressed at very low concentrations (e.g., transcription factors, cytokines, single-copy gene products) or proteins labeled with low energy particle-emitting radioisotopes, such as [^3H], include indirect autoradiography which utilizes intensifying screens for signal enhancement and fluorography.

Autoradiography is a photographic method for documenting the spatial distribution of radioisotopes within a given tissue, organism, cell organelle, or molecule. In autoradiography, the dried polyacrylamide gels bearing the radiolabeled proteins are placed in direct contact with the appropriate X-ray film where radioactive emissions react with the silver halides in the film emulsion, resulting in the formation of elemental silver atoms that can be distinctly visualized after the photographic development of the films. A radioactively labeled specimen, such as a slice of tissue or polyacrylamide gel, is placed in direct contact with a photographic emulsion intended for radiography in this technique. In the sample, the radioactive atoms will decay, and released radiation will cause individual silver halide grains in the emulsion, making them vulnerable to a photographic developer's conversion into metallic silver.

19.7.2 Chemiluminescence

Chemiluminescence takes place when light is emitted by a chemical reagent that contains stored energy. The reagent is typically stable and does not emit light, but can, for example, be transformed into a light-emitting product after contact with a particular enzyme. The enzyme horseradish peroxidase (HRP) conjugated to a secondary antibody is the catalyst that fulfills this role in most modern ECL systems. The light emitted is proportional to the amount of labeled compound in the sample and can be detected using CCD camera-based imagers as well as on X-ray films. As these antibody-based systems are commonly designed to target particular biomolecules, ECL is more robust than general colorimetric methods. Moreover, the technique is fast and sensitive; in seconds, signals are produced, and relatively small amounts of antigens and antibodies are typically consumed. As light is produced without an external source of excitation, there is no possibility of samples being photodamaged.

Chemiluminescence assays have lower background signal compared to absorbance and fluorescence assays that lead to greater sensitivity. Majority chemiluminescence-based reactions, however, possess low quantum efficiency and thus generate poor luminescence. This can limit the analytical assay applications.

Due to their excellent optical, electronic, and catalytic properties, the introduction of nanomaterials has introduced new capabilities into chemiluminescence assays in recent years; this could boost the efficiency of chemiluminescence assays as playing the role of catalyzers and fluorescence emission acceptors [15]. Compared to photoluminescence, chemiluminescence remains, despite the low quantum efficiency, which is an attractive choice for chemical analysis. This is due to three factors, namely, (1) improved signal-to-background and signal-to-noise ratios because of absence of a source of excitation; (2) inexpensive and stable instruments for assay; and (3) higher level of selectivity available. Recent developments such as the amino acid microsequencing and mass spectral analysis have made the effective identification of previously unidentified proteins contributing to a better understanding of the pathogenesis of diseases [16].

19.8 Mass Spectrometry Methods of Host-Pathogen Interaction Study

Mass spectrometry (MS) is crucial for the analysis of **proteomes** and is the method of choice for identifying proteins in any biological systems; hence, MS-based analysis platforms using several methodologies have been developed for the analysis of proteomes.

Mass spectrometry is the commonest of all the methods used for detection of analytes in proteomics and metabolomic research. However, the data analysis is very complex due to the huge amount of generated data that could be crunched by bioinformatics support and stand-alone proteomic software.

Proteins are characterized by MS analysis by intact proteins (top-down approach) or enzymatically digested protein **peptides** (bottom-up approach). In this method, the ions are created from neutral proteins, peptides, or metabolites, which are then separated according to their mass-to-charge ratio (m/z) and are detected to create a mass spectrum, characteristic of the molecular mass and/or structure [17]. In addition to that, several procedures, with or without stable isotope labeling, are used for protein quantitation (e.g., characterize changes in protein abundances between given biological states). Figure 19.5 shows a general flow chart of mass spectrometry.

Each analytical technique has its own different advantages and limitations in terms of many characteristics such as the instrument sensitivity, resolution, mass accuracy, dynamic range, and throughput.

Several techniques to determine quantitative analysis, including DIGE and ICAT labeling, can be coupled with tandem mass spectrometry. In order to study the metabolomics, other analytical platforms, such as nuclear magnetic resonance (NMR) spectroscopy and infrared spectroscopy, are used for metabolite identification [18]. The major advantages and disadvantages of the techniques used in proteomic studies of host-pathogen interactions are listed in Table 18.3.

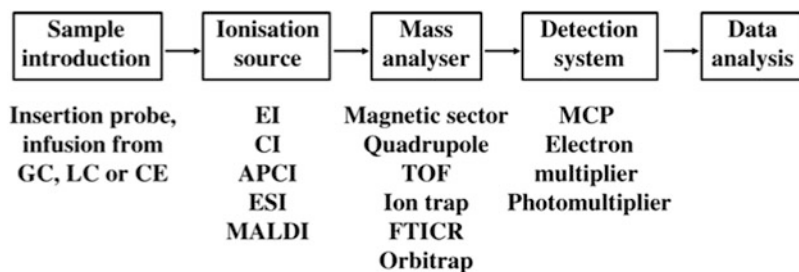


Fig. 19.5 General flow diagram for mass spectrometry. *APCI* atmospheric pressure chemical ionization, *CE* capillary electrophoresis, *CI* chemical ionization, *EI* electron impact ionization, *ESI* electrospray ionization, *FTICR* Fourier transform ion cyclotron resonance, *LC* liquid chromatography, *MALDI* matrix-assisted laser desorption/ionization, *MCP* microchannel ionization, *ToF* time of flight

19.9 Shotgun Proteomics

Shotgun proteomics is named for its similarity to shotgun sequencing, a method used in genomic sequencing, which has since been replaced by next-generation technologies. Given the rapid advancement in technology for molecular biology, shotgun proteomics has been there since a decade.

In shotgun proteomics, first, the proteins are extracted from a biological sample and then digested, following which a reverse phase liquid chromatography or electrophoresis can be used to fractionate the resulting peptides. Next, tandem mass spectrometry (MS) is performed on the sample, and the results are matched to previously known peptides.

While this high-throughput approach enables researchers to analyze multiple proteins that may be present in a sample, confusion can arise if the resulting peptides end up matching ambiguously to multiple peptides in the database. This can happen in especially complex samples, such as infectious disease samples that can contain proteins from multiple organisms. In order to circumvent this issue, a number of analysis methods and statistical algorithms have been used to confidently validate and reconstruct proteins. The majority of proteins detected are currently inaccessible due to limitations with traditional mass spectrometry, and strategies are being developed to more fully understand how to maximize MS coverage [19].

19.10 Gel-Free Proteomics

Gel-free proteomics eliminate few of the major experimental errors observed in case of gel-based approaches. This method also eliminates the extraction of protein from the gel base; hence the concentration of proteins/peptides for analysis will be in a detectable level. Also, the gel-free techniques can be well utilized for proteome

Table 18.3 Major advantages and disadvantages of various important proteomic techniques

Technique	Advantages	Disadvantages
2D-PAGE	Robust proteins can be isolated in pure form Posttranslational modifications analysis Study differential protein expression between types of cells Semi-quantitative protein expression assessment MS recognizes peptides derived predominantly from a single protein; more direct protein recognition	More manual work Problematic for very low or very high molecular weight proteins Salt ions
Protein microarray	High performance Specific to protein interactions with several molecular categories	Known proteins detected. Isolate protein in native conformation
Mass spectroscopy	High sensitivity Diversified automation This technique can be coupled with other proteomic platforms	MS recognizes peptides; final review of software groups peptides belonging to a protein; less direct protein identification The need for manual labor is low (<10%) rate of protein recognition No match with MR and pH experimentally measured
SELDI-TOF-MS “Surface-enhanced laser desorption ionization”	High sensitivity Clinical sample analysis, inbuilt automation, and little manual work	Needs clinical diagnostic validity Relatively costly A collection process prior to spectrum data review that has low quality
MALDI-TOF-MS “Matrix-assisted laser desorption ionization”	High sensitivity Relatively quick analysis	Needs comparatively pure samples More confusion in identifiers due to the lack of real sequence reliance on recognition It is important to have the same or high homology protein sequence in the database

quantification studies and comparison studies which play a significant part in analyzing the host-pathogen interaction process.

There are two methods of “gel-free approaches,” i.e., label-based approaches and label-free approaches [20]. These methods utilize the proficient liquid chromatography and mass spectrometry tools. This is a simple technique that follows sequential processes where proteins are isolated, digested (labeled/non-labeled), eluted on liquid chromatography, and detected and analyzed by mass spectrometry.

In this strategy, the protein will be enzymatically digested and subjected to high resolution chromatographic separation. The eluted peptides are then transferred to MS where m/z ratios are analyzed and chromatogram depicting signal intensities. The received signal intensities of peptides provide the direct measure identifying its

abundance in the sample [21]. The peptides are further fragmented by triple quadrupole mass analyzer, referred to as tandem MS (MS/MS). The data obtained from MS/MS provides the identity of proteins through comparison of peptide masses. The peptide masses of the protein are analyzed using established proteomics databases using softwares such as “Sequest” [13].

19.11 Labeling in Proteomic Study

Label-based approaches use specialized isotope/isobar tags that label proteins and peptides chemically or metabolically or enzymatically [22]. The separation of these labeled peptides are carried out using mass spectrometry. The label-based approaches identify unknown proteins through automation and multiplexing abilities. Labeled protein/peptides introduce mass shifts that distinguish the relative intensities of the proteins in the sample [23].

The most important labeling method (Fig. 19.6) used in the study of host-pathogen interaction is chemical labeling; this method uses chemically synthesized tags that will incorporate variable isotopes and isobars; this in turn will introduce mass difference within the labeled proteins (ICAT) and peptides (ICPL, iTRAQ, TMT) for their differential expression studies based on the abundance of peptides detected by their peak intensities (based on m/z) and MS/MS fragmentation. This method offers more accuracy and simultaneous comparison of more samples at the sample time, and comparison can also be made to control and treatments. This technique can be further classified into isotopic (ICAT and ICPL) and isobaric (iTRAQ and tandem mass tag (TMT) labeling) [24].

ICAT: Isotope-Coded Affinity Tags:

The tag consists of three functional elements i.e. iodoacetyl group (yellow) binds to thiol-specific groups, linker (blue) introduces mass shifts and biotin (red) used for reducing complexity by affinity purification.

ICPL: Isotope Coded Protein Labelling

Modified version of ICAT that permits multiplexing and labelling of almost all peptides.

iTRAQ: Isobaric Tags for Relative and Absolute Quantification

The tag consists of reporter (red) that introduces mass differences, balance group (blue) that maintains the similar weight of all reporter group in tags and NHS group (yellow) that binds specifically to peptides.

TMT: Tandem Mass Tag

This technique differs from iTRAQ with respect to the presence of additional linker group (white) and isobars used.

Fig. 19.6 Labeling methods used to study host-pathogen proteomics

19.12 Isobaric Mass Tagging in Mass Spectrometry

One of the important techniques in quantitative proteomics is the isobaric labeling; this is a mass spectrometry strategy where peptides or proteins are labeled using various chemical groups that have identical masses (isobaric) but will have varied distribution of heavy isotopes around their structure. An example of this technique for use in studying two keystone oral pathogens and their interaction with the host is explained. Further to the sample preparation, a tandem mass tag (TMT) isobaric mass tagging was done for labeling the samples. Two sets of labeling are done, namely, the TMT labels 128 and 130, where the phorbol myristate acetate (PMA)-treated protein digests will be labeled with TMT labels 127 and 130. Equal amounts of the labeled control and PMA-treated protein digests will be combined for mass spectrometry (MS) analysis [25].

In one of our study using such TMT labeling, a comparative analysis of several *Filifactor alocis* isolates showed heterogeneity in virulence modulation and its virulence potential. *F. alocis* which is one of the keystone pathogens can have possible interaction with other important periodontal pathogens such as the *Porphyromonas gingivalis*. This study envisages to explore the host-pathogen interaction and strain variation and host modulation; hence, in coculture with *P. gingivalis*, these *F. alocis* strains showed variations in their capacity for invasion of epithelial cells. In the oral microbiome study there are possibilities of synergistic interactions during polymicrobial infections that have resulted in enhanced pathogenesis of periodontopathogens such as *P. gingivalis*. This quantitative proteomics approach determined whether there is a similar mechanism(s) for *F. alocis* potentiating other oral microbes. It is likely that surface and secretory proteins from *F. alocis* play a role in this process.

In our overall results, the proteome profile of epithelial cells coinfecting with *F. alocis* and *P. gingivalis* strains showed activation of several eukaryotic proteins and pathway proteins that are involved in major cellular processes such as the inflammatory response, cell signaling, and cell death. The global proteome analysis of these hosts showed modulation in expression of 209 proteins [25]. *F. alocis* and *P. gingivalis* are important members of a complex multispecies biofilm that occupies the gingival crevice. Multiple interbacterial interactions are required for developing and maintaining the subgingival microbial community. Our study showed that the impact of these interspecies interactions on the host is very significant for their survival and marks their characteristic variations in the disease pattern. This study also has shown the inherent ability of specific factors from *F. alocis* to modulate multiple changes in the host cell proteome [25]. The identified unique molecular variations are responsible for the functional changes required to mediate the pathogenic process. The relative significance of specific *F. alocis* virulence factors can trigger the key host response and hence may cause a varied disease pathology.

Another approach to study host-pathogen infection is through usage of chemically reactive fluorescent dyes—and the technique called the “reactive probe-based chemical proteomic strategy” is one among the new techniques used for studying the host-pathogen interaction that is different from the previously developed methods. In

this study, the infected macrophages with *Salmonella typhimurium* are labeled with commonly used dyes such as the fluorescein. The study found un-important proteins in the involvement of host-pathogen interaction and virulence modulation [26].

19.13 Future of Proteomics in Pathogen Interaction Study

The proteome is a complex and dynamic entity that is always considered as an important source of data in studying the host-pathogen interaction and at the molecular level is characterized in terms of the sequence, structure, abundance, localization, modification, interaction, and biochemical function,. Hence, as the proteome is diverse, their technology or analysis and interpretation also stand diversified. Proteomics provides a robust and a global picture on the functioning cell in every stage of infection. The future of proteomics lies in the proteomics approaches that may be integrated with unique technologies seeking various scientific questions in order to study the interactions among both host and the pathogen side. One among the amalgamated approaches is the systems biology approach where several research have now focused on integrating the proteomics with metabolomics through bioinformatic analysis as a possible connector to identify the overall changes that happen in a system. In biomedical and life science studies, studying proteomes through functional consequences of genetic variation at different stages of a process advance our understanding of the eukaryotic proteome. Hence, we need to be monitoring simultaneously the variations in the genome structure, chromatin configuration (epigenomics), and gene expression (transcriptomics), in the prokaryotic proteome through protein expression. Such molecular interaction studies will pave way to identify novel drug targets through “interacto-proteome” designates in a specific system or process. Integrated genomics and proteomics monitor protein expression in targeted proteome that will enable further to characterize the interactions of the organism at the molecular level.

Despite the increase in value of proteomic study from human biomedical research throughout the life sciences, there has been a steady increase in using an integrative proteomics approach to study the prokaryotic systems. For example, studying the genome sequencing of epidemic and non-epidemic strains of the bacteria following an outbreaks will confirm localization of surface layer proteins. Recent techniques that amalgamate the proteomic analyses of prokaryotic cell wall proteins in tandem with genome sequencing will aid in bacterial genome annotation of genes and their functions. Such studies will provide candidate prophylaxis in later days to come.

Furthermore, mass spectrometry methods have been now used to study the glycoprotein moieties on the cell surface of pathogens which interact with the host cell. They are helpful in monitoring and detecting these important molecular biological changes. In addition to only studying the proteome, and in order to decipher the role of protein combinations in virulence of bacteria, a new glycol-proteomics method, termed glycan reductive isotope labeling (GRIL), that will identify the free glycans by reductive amination with the differentially coded stable

isotope tags is now being used to study the role of pathogen surface glycoprotein host interaction [27].

In recent trend, the “omics” technologies combine multidisciplinary expertise in biology, namely, the fields of genetics, computer science, and engineering and technology. This field of study throws new light in studying protein expression in correlation with protein functions. Hence, this methodology gives an overall picture of diseases especially the ones caused by variants (e.g., COVID-19) and thus could offer new insights into future treatment and prophylaxis.

19.14 Conclusion

In the current post-genomic era, proteomics is one among the important core technologies. This technique coupled with systems biology approaches is now used to understand molecular mechanisms underlying normal and disease states. Especially their progression and identification of critical diagnostic and prognostic biomarkers. Hence, proteomics plays an important role in tracing the various molecular modifications happening during the course of host-pathogen interaction. Hence, [proteomics](#) in addition to identifying the proteins present in a sample also can assess protein abundance and localization and identify modifications such as the [posttranslational modifications](#), isoforms, and molecular interaction. Hence, proteomics is indispensable to understanding the complexity of the interactive process between the host and pathogen. Many tools in improving protein analysis have become much important to understand and explore the mechanisms of action of bioactive molecules, find disease (bio)markers, and characterize new classes of drugs and pharmaceuticals such as the antibodies. Recent developments in protein analysis promote research beyond boundary in different fields of science. Moreover, advanced analytical tools could open up new possibilities in fields beyond protein science, venturing into new fields of polymer and biopharmaceutical research. Because, this field of protein study quenches the analytical challenge underlying voluminous big data that play an important role in important implications of health and disease.

References

1. Figeys D (2004) Combining different “Omics” technologies to map and validate protein-protein interaction in humans. *Brief Funct Genomic Proteomic* 2(4):357–336
2. Karr TL (2008) Application of proteomics to ecology and population biology. *Heredity* 100:200–206
3. Anderson NG, Anderson NL (1996) Twenty years of two-dimensional electrophoresis: past, present and future. *Electrophoresis* 17:443453
4. Biron DG, Moura H, Marché L, Hughes AL, Thomas F (2005a) Towards a new conceptual approach to “Parasitoproteomics”. *Trends Parasitol* 21:162168
5. Stebbins CE, Galan JE (2001) Structural mimicry in bacterial virulence. *Nature* 412 (6848):701–705

6. Covert BA, Spencers JS, Orme IM (2001) The application of proteomics in defining the T-cell antigens of *Mycobacterium tuberculosis*. *Proteomics* 1(4):574–586
7. Theodorescu D, Mischak H (2007) Mass spectrometry based proteomics in urine biomarker discovery. *World J Urol* 25:435–443. <https://doi.org/10.1007/s00345-007-0206-3>
8. Petricoin E, Zoon K, Kohn E, Barrett J, Liotta L (2002) Clinical proteomics: translating benchside promise into bedside reality. *Nat Rev* 1:683–695. <https://doi.org/10.1038/nrd891>
9. Vlahou A, Fountoulakis M (2005) Proteomic approaches in the search for disease biomarkers. *J Chromatogr B Analyt Technol Biomed Life Sci* 814:11–19. <https://doi.org/10.1016/j.jchromb.2004.10.024>
10. Bischoff R, Luidert TM (2004) Methodological advances in the discovery of protein and peptide disease markers. *J Chromatogr B* 803:2740
11. O'Farrell PH (1975) High resolution two dimensional electrophoresis of proteins. *J Biol Chem* 250(10):4007–4021
12. Nachimuthu Saraswathy, Ponnusamy Ramalingam, in Chapter [concepts and techniques in genomics and proteomics](#), 2011. Book: [Methods in Microbiology](#), 2006. Elsevier Publication
13. Chapman JD, Goodlett DR, Masselon CD (2014) Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom Rev* 33:452–470. <https://doi.org/10.1002/mas.21400>
14. Aruni AW, Roy F, Sandberg L, Fletcher HM (2012) Proteome variation among *Filifactor alocis* strains. *Proteomics* 12(22):3343–3364. <https://doi.org/10.1002/pmic.201200211>
15. Xu H, Liu B, Meng Y (2017) Ultrasensitive chemiluminescence assay for the lung cancer biomarker cytokeratin 21-1 via a dual amplification scheme based on the use of encoded gold nanoparticles and a toehold-mediated strand displacement reaction. *Microchim Acta* 184:3953–3959
16. Velde GV, Kucharíková S, Schrevens S, Himmelreich U, Van Dijck P (2013) Towards non-invasive monitoring of pathogen–host interactions during *Candida albicans* biofilm formation using *in vivo* bioluminescence. *Cell Microbiol* 16(1):115–130. <https://doi.org/10.1111/cmi.12184>
17. Brown M, Dunn WB, Dobson P, Patel Y, Winder CL, Francis-McIntyre S et al (2009) Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst* 134:1322–1332. <https://doi.org/10.1039/b901179j>
18. Rifai N, Gillette MA, Carr SA (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* 24:971–983. <https://doi.org/10.1038/nbt1235>
19. Marcotte EM (2007) How do shotgun proteomics algorithms identify proteins? *Nat Biotechnol* 25(7):755–757
20. Syahir A, Usui K, Tomizaki K, Kajikawa K, Mihara H (2015) Label and label-free detection techniques for protein microarrays. *Microarrays* 4:228–244
21. Chelius D, Bondarenko PV (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res* 1(4):317–323
22. Bantscheff M, Lemeer S, Savitski MM, Kuster B (2012) Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* 404(4):939–965. <https://doi.org/10.1007/s00216-012-6203-4>
23. Cosette Abdallah, Eliane Dumas-Gaudot, Jenny Renaut, Kjell Sergeant. Gel-Based and gel-free quantitative proteomics approaches at a glance. Volume 2012 |Article ID 494572 | <https://doi.org/10.1155/2012/494572>
24. Schulze WX, Usadel B (2010) Quantitation in mass-spectrometry-based proteomics. *Annu Rev Plant Biol* 61:491–516
25. Wilson Aruni A, Zhang K, Dou Y, Fletcher H (2014) Proteome analysis of coinfection of epithelial cells with *Filifactor alocis* and *Porphyromonas gingivalis* shows modulation of pathogen and host regulatory pathways. *Infect Immun* 82(8):3261–3274. <https://doi.org/10.1128/IAI.01727-14>
26. Ge J, Yao SQ (2017) Chemical proteomics of host pathogen interaction. *Chem Biol* 22(4):434–435. <https://doi.org/10.1016/j.chembiol.2015.04.002>

-
27. Xia B, Feasley CL, Sachdev GP, Smith DF, Cummings RD (2009) Glycan Reductive Isotope Labeling (GRIL) for quantitative glycomics. *Anal Biochem* 387(2):162–170. <https://doi.org/10.1016/j.ab.2009.01.028>



Significance of Post-translational Modifications in Apicomplexan Parasites

20

Priya Gupta, Rashmita Bishi, Sumbul Khan, Avi Rana, Nirpendra Singh, and Inderjeet Kaur

Abstract

Post-translational modifications (PTM) are the covalent modifications of amino acid constituents in cellular proteins. PTMs render complexity in proteome and are important regulators of protein functions. Apicomplexa is a group of obligate intracellular parasites that cause several human diseases including malaria and toxoplasmosis. PTMs have emerged out to play significant role in regulating gene expression in these parasites as well and have been described to regulate crucial parasite-specific molecular processes. Moreover, the enzymes that catalyze post-translational modifications have also been explored for therapeutic potential against these parasites. Due to their low stoichiometry, PTM-modified proteins are difficult to study. However, advancements in modern proteomics and mass spectrometry technologies offer solutions to investigate PTMs on cellular proteins. In this article, we aim to discuss about the status and significance of

P. Gupta · R. Bishi · S. Khan

Malaria Biology Group, International Centre for Genetic Engineering and Biotechnology, Aruna Asaf Ali Marg, New Delhi, India

A. Rana

Department of Biotechnology, Central University of Haryana, Mahendergarh, Haryana, India

N. Singh

Advanced Technology Platform Centre, Regional Centre for Biotechnology, Faridabad, Haryana, India

I. Kaur (✉)

Malaria Biology Group, International Centre for Genetic Engineering and Biotechnology, Aruna Asaf Ali Marg, New Delhi, India

Department of Biotechnology, Central University of Haryana, Mahendergarh, Haryana, India

e-mail: inderjeet@cuh.ac.in

post-translational modifications in apicomplexan parasites with special emphasis on phosphorylation and methylation.

Keywords

Post translational modifications · Apicomplexa · Plasmodium · Mass spectrometry

20.1 Introduction

Post-translational modifications (PTMs) refer to the covalent alterations of the polypeptide chain occurring in a cell following protein biosynthesis. Specific enzymes such as kinases, phosphatases, methyltransferases, demethylases, and acetyltransferases catalyze the process of post-translational modification of proteins in a cell. More than 300 different kinds of PTMs have been identified on cellular proteins involving as many as 15 amino acids [1]. These modifications include phosphorylation, methylation, acetylation, glycosylation, nitrosylation, ubiquitination, lipidation, and sumoylation. Proteolytic processing of proteins to obtain functional proteins is also considered to be a type of post-translational modification. PTMs of proteins increase proteome diversity in a cell and offer potent mechanism to rapidly and reversibly regulate protein function by changing structure, stability, localization, protein-protein interaction, etc., thereby affecting almost all aspects of normal cell biology. Post-transnationally modified proteins are, however, represented in sub-stoichiometric levels as compared to their unmodified counterparts making it extremely difficult to investigate PTM levels in the cellular milieu. Modern proteomics approaches coupled with high-resolution mass spectrometry offer solution to this problem by incorporating improved enrichment processes and sensitive methodology.

Post-translational modifications have emerged out to be the powerful regulators of cellular mechanisms of several pathogens responsible for causing severe diseases in humans. Apicomplexa consisting of a large group of protozoans is among such pathogenic organisms. Apicomplexans constitute a large phylum of protists, most of which are obligate intracellular parasites, several of which cause various infections in humans and animals. They are characterized by the presence of a special organ called apicoplast, a conserved set of specialized apical organelles called the rhoptries and micronemes—and a common mechanism of actin-based motility that is essential for invasion into their host cells. This group includes *Plasmodium* spp. (malaria), *Babesia* spp. (babesiosis), *Toxoplasma gondii* (toxoplasmosis), *Cryptosporidium parvum* (cryptosporidiosis), and *Cyclospora cayetanensis* (cyclosporiasis) [2]. Apicomplexans have a complex life cycle, and their distinctive biology represents deviation from the common “typical” eukaryotic behavior. *Plasmodium falciparum* and *Toxoplasma gondii* are the most prominent representatives of apicomplexan parasites responsible for causing severe diseases worldwide. Malaria accounts for more than 200 million annual infections with nearly 445,000 deaths

globally, and *Plasmodium falciparum* is responsible for the majority of malaria-related casualties worldwide [3]. The life cycle of *Plasmodium falciparum* is highly complex and alternates between two hosts: human and *Anopheles* mosquito involving several morphologically distinct stages both in humans and in the mosquito vector [4]. Although transcription regulation plays a very significant role in helping parasite to adapt to distinct environments in different cell types in two hosts, emerging evidences suggest that post-translational modifications play an important role in regulation of fundamental processes like host invasion, cell proliferation and differentiation, cell signaling, protein-protein interaction, and epigenetic control of gene expression [5]. *Toxoplasma gondii* is an obligate intracellular parasite infecting approximately 30%–50% of the human population worldwide. The symptoms of the disease are mild in healthy humans; however, the disease can be severe and sometimes fatal in children and immune-compromised individuals [6]. Both *Plasmodium falciparum* and *Toxoplasma gondii* use an actin-myosin-based motility system to actively invade host cells, a characteristic of apicomplexan organisms.

During the last decade, the role of post-translational modifications has emerged out to be a powerful mechanism of regulation of parasitic proteins [5, 7–11] and has been implicated in crucial parasitic processes like gliding motility, invasion, egress, and var gene expression. Several PTMs have been studied in apicomplexan parasites, but here, we will focus on phosphorylation and methylation. Proteolytic processing of proteins as a mode of regulating their activity is an exceedingly dominant feature of apicomplexan biology, which will also be discussed here.

20.2 Phosphorylation

Phosphorylation is one of the widely studied post-translational modifications of cellular proteins. It is the reversible addition of phosphoryl moiety on serine, threonine, and tyrosine amino acids, and the reaction is catalyzed by special enzymes called kinases. Phosphorylation is a ubiquitous modification regulating almost every essential cellular process. The status of phosphorylation of a protein in a cell is regulated by the interplay of two enzymes: kinases and phosphatases, the former catalyzes the addition of phosphate group while the latter catalyzes the removal of the same from a given protein substrate. Human genome encodes more than 500 kinases and ~200 phosphatases [12], regulating more than 200,000 phosphorylation sites in the human proteome.

In Apicomplexa too, phosphorylation has got its attention from the researchers all over the world and turned out to be the most extensively studied post-translational modification of parasitic proteins. *Plasmodium* genome encodes approximately 99 protein kinases (PKs), and phylogenetic studies show that it has almost all the eukaryotic PKs except the tyrosine protein kinases [13, 14]. The “kinome” of *Plasmodium falciparum* has been well described in a study by Ward et al. [14], which identified a novel kinase family, FIKK, having 20 members in *Plasmodium falciparum*. FIKK is an Apicomplexa-specific kinase family, and most of the apicomplexans possess one kinase from this family. Protein phosphorylation has

been implicated in almost all the essential processes of the malaria parasite including all stages of its life cycle. Most importantly, the essential role of protein phosphorylation has been established in gliding motility, invasion, and egress of *Plasmodium* [11]. A study described one of the first global phosphoproteomes of *P. falciparum* at asexual blood stages and by using reverse-genetics approach, identified 36 parasite kinases essential for the survival of the parasite [15]. A kinome-wide knockout strategy was used to investigate the role of individual kinases in parasite development in the rodent parasite *P. berghei* [16]. Several other studies independently established the role of protein phosphorylation in *P. falciparum* development and life cycle [17–20]. Stage-specific global quantitative phosphoproteomic changes were also established at ring, trophozoite, and schizont stages of the *P. falciparum* revealing the changes occurring in the levels of phosphorylation in parasitic proteins when the parasite switches from one stage of its life cycle to another [19]. Two kinases, PfPKA and PfPKG, which are regulated by second messenger cyclic nucleotide, were shown to perform an essential role in the phosphorylation cascades involved in the processes of invasion and egress of host cell [21, 22].

The phylogenomic analysis revealed that *T. gondii* encodes 108 protein kinase and 51 pseudokinase genes [23, 24]. The analysis further divulged that the parasitic protein kinases possess major CMGC and CAMK groups but lack tyrosine kinases or receptor guanylate cyclases. Majority of apicomplexan-specific *T. gondii* kinases share orthologues with *Plasmodium* [23, 24]. Out of 108 active protein kinases, 78 do not have orthologue in humans. Majority of *T. gondii* kinases differ in structure and function from mammalian protein kinases as 55% of them are species-specific while 15% are secretory, such as rhoptry kinases [23]. Using modern proteomics and high-resolution mass spectrometry, a study compared the global phosphoproteomes of two members of Apicomplexa, *P. falciparum* and *T. gondii*, and indicated towards the evolutionary mechanisms associated with phosphorylation motifs in these two parasites [17]. *Cryptosporidium parvum* kinome consists of 73 protein kinases identified with intact catalytic triads including members of AGC, atypical, CaMK, CK1, CMGC, and TKL groups [25]. Most of these kinases are specific to the parasite as 25% of the total identified kinases do not have orthologue outside of *Cryptosporidium* spp. [25].

Apicomplexa are characterized by the presence of a plant like family of calcium-dependent kinases (CDPKs). *Plasmodium falciparum* genome encodes seven such calcium-dependent protein kinases (PfCDPK1–7); *Plasmodium berghei* has six CDPK homologues, while in *T. gondii*, 14 such genes have been identified [26]. CDPKs are important regulators and mediators of calcium signaling in these parasites. The fact that CDPKs have been found to be crucial for parasite development and are not found in mammals, makes CDPKs the ideal and interesting targets for developing therapeutic interventions. The essential role of different CDPKs in malaria parasite development has been summarized elsewhere [26]. Different CDPKs have been implicated in different parasitic processes and at different life stages. PfCDPK1 is indispensable in *P. falciparum* as evidenced by gene knockout studies and has been shown to play crucial roles in essential parasitic processes like invasion, gliding motility, and egress. PfCDPK1 was found to be important for

microneme discharge and invasion of merozoites into erythrocytes as inhibitors of PfCDPK1 blocked these two processes [27]. A very recent study investigated the phosphorylation-dependent differentially regulated signaling pathways during invasion process of *P. falciparum* [28]. Using global quantitative phosphoproteomics, the authors found changes in 143 Ca²⁺-dependent protein phosphorylation sites during invasion and deciphered a phosphorylation-dependent multi-protein complex involving multiple kinases including PfCDPK1. This complex is crucial for the process of invasion into host cells. PfCDPK1 also phosphorylates *Plasmodium falciparum* serine repeat antigen 5 (PfSERA5), a protease essential in egress of merozoites out of erythrocytes implicating vital function of PfCDPK1 in egress pathway [29]. PfCDPK1 was also shown to phosphorylate components of motor complex in parasite [30]. These findings bring forth the highly important role played by PfCDPK1 in *P. falciparum* biology and PfCDPK1, itself, is a heavily phosphorylated kinase. TgCDPK3, the orthologue of PfCDPK1 in *T. gondii*, regulates the parasite egression process. TgCDPK3 phosphorylates myosin A, which is essential for initiation of motility and egress providing a mechanistic link between TgCDPK3 regulation and the lytic cycle [31].

Recently, the role of PfCDPK2 was found crucial for male gametocyte exflagellation and ookinete development [32]. The authors envisage that PfCDPK2 may not have essential role at asexual blood stages. However, PfCDPK2 was also found to be having three phosphorylation sites in vivo. CDPK4 has been found to be important for the process of exflagellation, both in *P. falciparum* and in *P. berghei*, as selective inhibitors of CDPK4 were able to block exflagellation in both parasites [33, 34]. Therefore, a critical role of CDPK4 has been established in sporozoite motility and malaria transmission [35]. The orthologue of CDPK4 in *T. gondii* is TgCDPK1 that is essential for egress and invasion pathways [36]. In *P. falciparum*, a genetic knockout study deciphered that PfCDPK5-deficient parasites were unable to egress out of host erythrocytes and were found to be growth stalled at mature schizont stages establishing a vital role of PfCDPK5 in parasite egress pathway [37]. A first-ever comprehensive genome-wide expression analysis of the CDPK gene family in *T. gondii* revealed multiple functions of the kinase members and showed that CDPK6 is involved in oocyst development [38]. An important role for PfCDPK7 in the erythrocytic asexual stages of the parasite was discovered [39]. It was found that PfCDPK7 is involved in phosphatidylinositol phosphate (PIP) signaling through its interaction with PI(4,5)P₂ via its pleckstrin homology domain. Genetic disruption of PfCDPK7 caused growth defects in the malaria parasite [39]. TgCDPK7 has a crucial function in early cell division and is necessary for precise maintenance of centrosome integrity during cell division [40]. The TgCDPK7-depleted parasites showed growth defects with impaired cell division, while other processes like motility, egress, and microneme discharge were unaffected [40]. Despite numerous studies, the comprehensive knowledge about the exact pathways and substrates of these calcium-dependent kinases is not fully understood in Apicomplexa which necessitates further investigations into CDPK-associated signaling mechanisms, leading to establish novel pathways of parasite biology. Recently, using gene knockout strategy followed by global quantitative

phosphoproteomics approaches, the signaling pathways of PfCDPK1 were deduced at asexual blood stages of malaria parasite [41]. However, many such studies need to be performed for individual CDPKs to obtain comprehensive knowledge about their signaling pathways in Apicomplexa to exploit them for their therapeutic potential.

The reversible status of phosphorylation on cellular proteins is maintained by interplay among kinases and phosphatases. A genome-wide in silico analysis described 67 phosphatases in *P. falciparum* [42]. The phylogenetic analysis revealed that *Plasmodium* phosphatases possess considerable proximity with apicomplexans having six *Plasmodium*-specific phosphatases. The authors found 33 putative phosphatases, which did not have orthologues in humans [42]. Protein phosphatases of *P. falciparum* possess low substrate specificity and a high catalytic activity [43]. Most of the eukaryotes have seven subfamilies of phosphoprotein phosphatases: PP1, PP2A, PP2B, PP4, PP5, PP6 and PP7. *Toxoplasma gondii* and *P. falciparum* possess all of these phosphatase subfamilies having one member in each subfamily [44]. However, *T. gondii* has two members of PP2A subfamily. Further, *C. parvum* lacks PP6 and PP7, while *B. bovis* lacks PP2B and PP6 [44]. Despite these differences, the phosphoprotein phosphatases of apicomplexans are highly conserved. Apicomplexans contain three families of phosphatases that are unique to them and not present in humans indicating towards the potential of using parasite protein phosphatases as drug targets.

20.3 Methylation

Methylation is defined as the addition of methyl group at the amino acid residue within a protein substrate. This process is catalyzed by enzymes called methyltransferases [10]. Methylation is added post-transnationally on arginine and lysine amino acids and is associated with gene expression. According to Swiss-Prot, methylation is termed as the fourth most abundant PTM [45].

Arginine methylation is an important post-translational modification, found in both nuclear and cytoplasmic proteins. It is an epigenetic regulator of a number of cellular processes including mRNA splicing, translation, cell signaling, cell death, and DNA damage repair [46, 47]. The methylation of specific histone residues modulates gene expression. The transfer of methyl group to specific arginine residue is carried out by a class of enzymes, viz., protein arginine methyltransferases (PRMTs) [46]. Different isoforms of PRMTs exist in different organisms having different cellular localization and also substrate specificity; however, they cross-regulate each other. Some motifs of PRMTs are phosphorylated or ubiquitinated for regulation of their own activity [48]. It has also been reported that PRMTs undergo oligomerization for their activation. Krause et al. [49] reported phylogenetic analysis of PRMTs, showing no homologue identification in bacteria and archaea suggesting emanation of PRMTs in early eukaryotic lineages.

Plasmodium genome encodes for three putative PRMT candidates, PRMT1, PRMT5, and CARM1 (PlasmoDB). PfPRMT1 has been characterized to be localized both in the cytoplasm and in the nucleus. It exhibits the methyltransferase

activity with extended N-terminal region; however, in the absence of this region, the activity is impaired [50]. The study also revealed that PfPRMT1, like its mammalian counterparts, does retain the capacity of oligomerization despite having very less sequence homology. The oligomerization is critical for the appropriate functioning of the enzyme [50]. PfPRMT1 could methylate histones (H4 and H2A) and other substrates exhibiting high-substrate turnover rates than mammalian PRMT1 [50]. The other two putative candidates are PfPRMT5 and PfCARM1, named so because of their sequence homologies [51]. However, PfCARM1 is relatively divergent from its homologues and is more closely related to the HsPRMT3. The putative role of PfPRMT5 has been suggested in a methylation-dependent assembly of spliceosomal core complex in the parasite, wherein PfPRMT5 is believed to be interacting with the components of this complex at blood stages [52].

Despite extensive research, it is indeed challenging to delineate the essential functions of various protein arginine methyltransferases and unravel their crucial role in the life cycle of the parasite. Due to the presence of only three PRMTs in *Plasmodium* genome, it is expected that these enzymes may have broader substrate specificity to perform vast variety of functions in the malaria parasite. Recently, Zeeshan et al. [53], revealed the significance of arginine methylation at asexual stages of *P. falciparum* using global proteomics approaches. The study used methyl-specific antibodies to enrich the methylated proteins from the ring, trophozoite, and schizont stages of the parasite and predicted that protein methylation of arginine residues is a widespread phenomenon in *Plasmodium*.

Toxoplasma gondii genome is believed to possess five isoforms of PRMTs (1–5), out of which two, TgPRMT1 (methylates H4R3) and TgCARM1 (methylates H3R17), have been validated to exhibit methyltransferase activities [51]. TgCARM1-mediated methylation has been implicated in gene regulation during parasite development. Two other methyltransferase genes from the parasite are homologous to type II HsPRMT5 and HsPRMT3, respectively. TgPRMT (GT1_073730) is highly unique and non-canonical because of its exceptionally large size, and it lacks significant homology to any other PRMT characterized so far [51]. TgPRMT1, the major methyltransferase, is concentrated in the apical region and maintains centrosomal stoichiometry, critical for cell division (during cytokinesis as well as karyokinesis) and for the formation of normal daughter buds [54]. TgPRMT1 plays an important role in RNA interference by methylating arginine residues on Tg argonaute N-terminal RGG domain and recruiting TSN, essential for cleavage of target RNA [55]. The different stages show differential subcellular location of TgPRMT5. In tachyzoite stage, TgPRMT5 localizes in the cytosol and shifts to the nucleus during the bradyzoite stage (cyst), suggesting that it plays an essential role in stage conversion in *Toxoplasma gondii* [56]. Small molecules such as AMI-1 exhibit anti-Arg N-methyltransferase activity by inhibiting the action of TgCARM1 but have minimal effect on TgPRMT1, whereas AMA-2 behaves antagonistic to AMI-1 with respect to CARM1 and agonistic to PRMT1 [57]. Recently, the extent and significance of arginine mono-methylation in *T. gondii* were established [58]. The study described that arginine monomethylated (MMA) proteins form almost 5% of the *T. gondii* proteome and many of them are heavily

modified at arginine residues having as many as seven MMA sites. The MMA proteins of *T. gondii* are enriched in DNA- and RNA-binding proteins are thought to be involved in a variety of parasitic functions [58].

The process of transferring methyl group on lysine residues is catalyzed by lysine methyltransferases (PKMTs). *Plasmodium falciparum* genome codes for ten histone lysine methyltransferases (HKMTs). HKMTs are a large family of at least ten histone methyltransferases (HKMT) containing SET domain and three histone lysine demethylases (HDMs) divided into two families: the lysine-specific demethylase 1 (LSD1) and JmjC domain-containing histone demethylases (JHDMs). H3K4, H3K39, H3K9, and H4K20 are methylated by well-described four HKMTs (PfSET1, PfSET2, PfSET3, PfSET8, respectively). One LSD1 (PflO57w) and two JHDMs (MAL8P1.111 and PFF0135w) are also encoded by *Plasmodium* genome. The specificity of various HKMTs and HDMs are not only limited to substrate but also for different methyl states (mono, di, tri) [59–62].

The extent and role of protein lysine methylation on malaria parasite proteins were recently described by a proteome-wide study [63]. Using methylated lysine-specific antibodies to enrich the lysine-methylated proteins followed by LC-MS/MS analysis, 605 lysine-methylated sites corresponding to 422 proteins at three different stages of *P. falciparum* were identified. The results indeed revealed that lysine methylation is an extensive modification in plasmodial proteins covering a wide range of proteins involved in diverse cellular functions such as transport, protein folding, nucleotide metabolic processes, homeostatic processes, and chromatin organization [63]. Despite the elucidation of the global status of protein methylation in *Plasmodium falciparum*, the functional significance of methyltransferases still remains elusive and demands for extensive research in this area.

The emerging drug resistance among parasite species necessitates research and development efforts targeted towards elucidation of novel parasite-specific pathways and establishing new drug candidate targets. Parasitic methyltransferases could be the ideal candidates to exploit them for novel therapeutic potential due to their less homology with their mammalian counterparts and their broad range of substrates having a deep impact on parasite biology. Moreover, the parasitic methyltransferases characterized so far exhibit different activity kinetics which could be exploited to develop small-molecule inhibitors against them.

20.4 Proteolytic Processing

Apicomplexans employ a number of proteolytic events during their life cycle to ensure parasite survival and pathogenesis [64]. The proteases act at different stages of the life cycle mediating numerous functions such as cell death, cell cycle progression, and cellular motility. A lot of proteins that participate in the process of invasion inside host cell are proteolytically cleaved when they are trafficked or exported to the parasite surface. A major class of proteases, subtilisin-like serine proteases, has important roles to play in the process of invasion. PFSUB1 is involved in the processing of MSP1/6/7 which is required for the invasion of parasite inside

the erythrocytes [65–67]. PfSUB2 cleaves the ectodomain of MSP1 and AMA1 [68, 69]. This processing is highly crucial for releasing the adhesion complexes for the parasite invasion. In toxoplasma, TgSUB1 is required for the processing of micronemal proteins like MIC2, MIC4, and M2AP. These proteins help the attachment of tachyzoites to the host cell. Another rhoptry protein, TgSUB2, is involved in the maturation of rhoptry proteins and is highly essential in the life cycle of *T. gondii* [70, 71]. Further, a rhomboid family of serine proteases is crucial for the protein shedding during the process of invasion [72, 73]. TgROM4 participates in the processing of adhesion proteins like MIC2, MIC3, and AMA1. In *plasmodium*, PfROM1 and PfROM4 are involved in the processing of different proteins that contain a transmembrane domain and are important in the process of parasite invasion [74]. PfROM4 has also been shown to be involved in the cleavage of erythrocyte binding like adhesins.

Once the parasite has invaded the host cell, different proteases play important roles in maintaining the growth and development of parasite inside the host cell. One of the processes that is highly instrumental for the survival of the malaria parasite is hemoglobin degradation. During the trophozoite stage, the parasite ingests hemoglobin from the host erythrocyte, which is utilized to generate amino acids for the protein synthesis in the parasite. The degradation of hemoglobin also generates space for the growth of the parasite and helps maintain the osmotic balance of the cell. Different cysteine and aspartic proteases are involved in the process of hemoglobin degradation. The process is initiated by aspartic proteases and plasmepsins [75–77]. Plasmepsins I and II initiate the catabolism of hemoglobin. This is further acted upon by the plasmepsin IV, which acts in the cleavage of peptides generated by plasmepsins I and II. Gene knockouts of the plasmepsins confer a growth defect on the parasite development. The peptides thus generated are further acted upon by a cysteine family of proteases called falcipains [78]. Finally, the aminopeptidases generate amino acids, which are finally used by the parasite for its protein synthesis.

Based on the presence of a variety of proteases and their equally diverse substrates, different classes of proteases are, therefore, important targets for the development of novel therapeutics.

20.5 Conclusion

Emerging evidences clearly indicate a highly significant and widespread role played by post-translational modifications in regulating parasite life cycles in Apicomplexa. Apart from the modifications discussed here, numerous others have been described and studied in apicomplexan parasites, further establishing the importance of post-translational regulation of gene expression in Apicomplexa. *Plasmodium* exhibits relatively rigid transcriptional machinery and possesses comparatively small number of transcriptional factors indicating towards another way of regulation of gene expression at post-transcriptional and post-translational levels allowing the parasites to adapt in different hosts and different cell types. The knowledge of parasitic processes controlled by post-translational protein modifications is still not sufficient

and demands further research and investigations into this field. Moreover, the enzymes that catalyze these PTMs could be explored for anti-parasitic drug discovery.

References

1. Uy R, Wold F (1977) Posttranslational covalent modification of proteins. *Science* 198:890–896
2. Lourido S, Moreno SN (2015) The calcium signaling toolkit of the Apicomplexan parasites *Toxoplasma gondii* and *Plasmodium* spp. *Cell Calcium* 57(3):186–193
3. WHO report (2017). <https://www.who.int/malaria/publications/world-malaria-report-2017/en/>
4. Sherman IW (2005) Molecular approaches to malaria. ASM, Washington
5. Doerig C, Rayner JC, Scherf A, Tobin AB (2015) Post-translational protein modifications in malaria parasites. *Nat Rev Microbiol* 13(3):160–172
6. Keeley A, Soldati D (2004) The glideosome: a molecular machine powering motility and host-cell invasion by Apicomplexa. *Trends Cell Biol* 14:528–532
7. Foth BJ, Zhang N, Mok S, Preiser PR, Bozdech Z (2008) Quantitative protein expression profiling reveals extensive post-transcriptional regulation and post-translational modifications in schizont-stage malaria parasites. *Genome Biol* 9:R177
8. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC et al (2004) Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Res* 14:2308–2318
9. Llinas M, Bozdech Z, Wong ED, Adai AT, DeRisi JL (2006) Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Res* 34:1166–1173
10. Chung DW, Ponts N, Cervantes S, Le Roch KG (2009) Post-translational modifications in Plasmodium: more than you think. *Mol Biochem Parasitol* 168(2):123–134
11. Yakubu RR, Weiss LM, Silmon de Monerri NC (2018) Post-translational modifications as key regulators of apicomplexan biology: insights from proteome-wide studies. *Mol Microbiol* 107(1):1–23
12. Sacco F, Perfetto L, Castagnoli L, Cesareni G (2012) The human phosphatase interactome: an intricate family portrait. *FEBS Lett* 586(17):2732–2739
13. Anamika, Srinivasan N, Krupa A (2005) A genomic perspective of protein kinases in *Plasmodium falciparum*. *Proteins* 58(1):180–189
14. Ward P, Equinet L, Packer J, Doerig C (2004) Protein kinases of the human malaria parasite *Plasmodium falciparum*: the kinome of a divergent eukaryote. *BMC Genomics* 5(1):79
15. Solyakov L et al (2011) Global kinomic and phospho-proteomic analyses of the human malaria parasite *Plasmodium falciparum*. *Nat Commun* 2:565
16. Tewari R et al (2010) The systematic functional analysis of Plasmodium protein kinases identifies essential regulators of mosquito transmission. *Cell Host Microbe* 8:377–387
17. Treeck M, Sanders JL, Elias JE, Boothroyd JC (2011) The phosphoproteomes of Plasmodium falciparum and *Toxoplasma gondii* reveal unusual adaptations within and beyond the parasites' boundaries. *Cell Host Microbe* 10:410–419
18. Lasonder E et al (2012) The *Plasmodium falciparum* schizont phosphoproteome reveals extensive phosphatidylinositol and cAMP-protein kinase A signaling. *J Proteome Res* 11:5323–5337
19. Pease BN et al (2013) Global analysis of protein expression and phosphorylation of three stages of *Plasmodium falciparum* intraerythrocytic development. *J Proteome Res* 12:4028–4045
20. Collins MO, Wright JC, Jones M, Rayner JC, Choudhary JS (2014) Confident and sensitive phosphoproteomics using combinations of collision induced dissociation and electron transfer dissociation. *J Proteome* 103:1–14

21. Leykauf K et al (2010) Protein kinase a dependent phosphorylation of apical membrane antigen 1 plays an important role in erythrocyte invasion by the malaria parasite. *PLoS Pathog* 6: e1000941
22. Collins CR et al (2013) Malaria parasite cGMP-dependent protein kinase regulates blood stage merozoite secretory organelle discharge and egress. *PLoS Pathog* 9:e1003344
23. Peixoto L, Chen F, Harb OS, Davis PH, Beiting DP, Brownback CS, Ouloguem D, Roos DS (2010) Integrative genomic approaches highlight a family of parasite specific kinases that regulate host responses. *Cell Host Microbe* 8:208–218
24. Wei F, Wang W, Liu Q (2013) Protein kinases of *Toxoplasma gondii*: functions and drug targets. *Parasitol Res* 112:2121–2129
25. Artz JD, Wernimont AK, Allali-Hassani A et al (2011) The *Cryptosporidium parvum* kinome. *BMC Genomics* 12:478
26. Ghartey-Kwansah G, Yin Q, Li Z, Gumper K, Sun Y, Yang R, Wang D, Jones O, Zhou X, Wang L, Bryant J, Ma J, Boampong JN, Xu X (2020) Calcium-dependent protein kinases in malaria parasite development and infection. *Cell Transplant* 29:963689719884888
27. Bansal A, Singh S, More KR, Hans D, Nangalia K, Yogavel M, Sharma A, Chitnis CE (2013) Characterization of *Plasmodium falciparum* calcium-dependent protein kinase 1 (PfCDPK1) and its role in microneme secretion during erythrocyte invasion. *J Biol Chem* 288 (3):1590–1602
28. More KR, Kaur I, Giai Gianetto Q, Invergo BM, Chaze T, Jain R, Huon C, Gutenbrunner P, Weisser H, Matondo M, Choudhary JS, Langsley G, Singh S, Chitnis CE (2020) Phosphorylation-dependent assembly of a 14-3-3 mediated signaling complex during red blood cell invasion by *Plasmodium falciparum* merozoites. *MBio* 11(4):e01287–20
29. Iyer GR, Singh S, Kaur I, Agarwal S, Siddiqui MA, Bansal A, Kumar G, Saini E, Paul G, Mohammed A, Chitnis CE, Malhotra P (2018) Calcium-dependent phosphorylation of *Plasmodium falciparum* serine repeat antigen 5 triggers merozoite egress. *J Biol Chem* 293 (25):9736–9746
30. Green JL, Rees-Channer RR, Howell SA, Martin SR, Knuepfer E, Taylor HM, Grainger M, Holder AA (2008) The motor complex of *Plasmodium falciparum*: phosphorylation by a calcium-dependent protein kinase. *J Biol Chem* 283(45):30980–30989
31. Garrison E, Trecek M, Ehret E, Butz H, Garbuz T, Oswald BP, Settles M, Boothroyd J, Arrizabalaga G (2012) A forward genetic screen reveals that calcium-dependent protein kinase 3 regulates egress in *Toxoplasma*. *PLoS Pathog* 8:e1003049
32. Bansal A, Molina-Cruz A, Brzostowski J, Mu J, Miller LH (2017) *Plasmodium falciparum* calcium-dependent protein kinase 2 is critical for male gametocyte exflagellation but not essential for asexual proliferation. *MBio* 8(5):e01656–e01617
33. Ojo KK, Pfander C, Mueller NR, Burstroem C, Larson ET, Bryan CM, Fox AM, Reid MC, Johnson SM, Murphy RC, Kennedy M, Mann H, Leibly DJ, Hewitt SN, Verlinde CL, Kappe S, Merritt EA, Maly DJ, Billker O, Van Voorhis WC (2012) Transmission of malaria to mosquitoes blocked by bumped kinase inhibitors. *J Clin Invest* 122:2301–2305
34. Ojo KK, Eastman RT, Vidadala R, Zhang Z, Rivas KL, Choi R, Lutz JD, Reid MC, Fox AM, Hulverson MA, Kennedy M, Isoherranen N, Kim LM, Comess KM, Kempf DJ, Verlinde CL, Su XZ, Kappe SH, Maly DJ, Fan E, Van Voorhis WC (2014) A specific inhibitor of PfCDPK4 blocks malaria transmission: chemical-genetic validation. *J Infect Dis* 209:275–284
35. Fang H, Klages N, Baechler B, Hillner E, Yu L, Pardo M, Choudhary J, Brochet M (2017) Multiple short windows of calcium-dependent protein kinase 4 activity coordinate distinct cell cycle events during *Plasmodium* gametogenesis. *eLife* 6:e26524
36. Lourido S, Shuman J, Zhang C, Shokat KM, Hui R, Sibley LD (2010) Calcium-dependent protein kinase 1 is an essential regulator of exocytosis in *Toxoplasma*. *Nature* 465 (7296):359–362
37. Dvorin JD, Martyn DC, Patel SD, Grimley JS, Collins CR, Hopp CS, Bright AT, Westenberger S, Winzeler E, Blackman MJ, Baker DA et al (2010) A plant-like kinase in

- Plasmodium falciparum* regulates parasite egress from erythrocytes. *Science* 328 (5980):910–912
38. Wang J, Huang S, Zhang N, Chen J, Zhu X (2015) Genome-wide expression patterns of calcium-dependent protein kinases in *Toxoplasma gondii*. *Parasit Vectors* 8:304
 39. Kumar P, Tripathi A, Ranjan R, Halbert J, Gilberger T, Doerig C, Sharma P (2014) Regulation of *Plasmodium falciparum* development by calcium-dependent protein kinase 7 (PfCDPK7). *J Biol Chem* 289(29):20386–20395
 40. Morlon-Guyot J, Berry L, Chen CT, Gubbels MJ, Lebrun M, Daher W (2014) The *Toxoplasma gondii* calcium-dependent protein kinase 7 is involved in early steps of parasite division and is crucial for parasite survival. *Cell Microbiol* 16(1):95–114
 41. Kumar S, Kumar M, Ekka R, Dvorin JD, Paul AS, Madugundu AK, Gilberger T, Gowda H, Duraisingh MT, Keshava Prasad TS, Sharma P (2017) PfCDPK1 mediated signaling in erythrocytic stages of *Plasmodium falciparum*. *Nat Commun* 8(1):63
 42. Pandey R, Mohammed A, Pierrot C, Khalife J, Malhotra P, Gupta D (2014) Genome wide in silico analysis of *Plasmodium falciparum* phosphatome. *BMC Genomics* 15:1024
 43. Wilkes JM, Doerig C (2008) The protein-phosphatome of the human malaria parasite *Plasmodium falciparum*. *BMC Genomics* 9:412
 44. Yang C, Arizabalaga G (2017) The serine/threonine phosphatases of apicomplexan parasites. *Mol Microbiol* 106(1):1–21
 45. Khoury GA, Baliban RC, Floudas CA (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep* 1:90
 46. Bedford MT (2007) Arginine methylation at a glance. *J Cell Sci* 120(Pt 24):4243–4246
 47. Zhang J, Jing L, Li M, He L, Guo Z (2019) Regulation of histone arginine methylation/demethylation by methylase and demethylase (Review). *Mol Med Rep* 19(5):3963–3971
 48. Guccione E, Richard S (2019) The regulation, functions and clinical relevance of arginine methylation. *Nat Rev Mol Cell Biol* 20:642–657
 49. Krause CD, Yang ZH, Kim YS, Lee JH, Cook JR, Pestka S (2007) Protein arginine methyltransferases: evolution and assessment of their pharmacological and therapeutic potential. *Pharmacol Ther* 113(1):50–87
 50. Fan Q, Miao J, Cui L, Cui L (2009) Characterization of PRMT1 from *Plasmodium falciparum*. *Biochem J* 421(1):107–118
 51. Fisk JC, Read LK (2011) Protein arginine methylation in parasitic protozoa. *Eukaryot Cell* 10(8):1013–1022
 52. Hossain M, Sharma S, Korde R et al (2013) Organization of *Plasmodium falciparum* spliceosomal core complex and role of arginine methylation in its assembly. *Malar J* 12:333
 53. Zeeshan M, Kaur I, Joy J, Saini E, Paul G, Kaushik A, Dabral S, Mohammed A, Gupta D, Malhotra P (2017) Proteomic identification and analysis of arginine-methylated proteins of *Plasmodium falciparum* at asexual blood stages. *J Proteome Res* 16(2):368–383
 54. El Bissati K, Suvorova ES, Xiao H, Lucas O, Upadhyaya R, Ma Y, Angeletti RH, White MW, Weiss LM, Kim K (2016) *Toxoplasma gondii* Arginine Methyltransferase 1 (PRMT1) is necessary for centrosome dynamics during Tachyzoite cell division. *MBio* 7(1):e02094–e02015
 55. Musiyenko A, Majumdar T, Andrews J, Adams B, Barik S (2012) PRMT1 methylates the single Argonaute of *Toxoplasma gondii* and is important for the recruitment of Tudor nuclease for target RNA cleavage by antisense guide RNA. *Cell Microbiol* 14(6):882–901
 56. Liu M, Li FX, Li CY, Li XC, Chen LF, Wu K, Yang PL, Lai ZF, Liu TK, Sullivan WJ Jr, Cui L, Chen XG (2019) Characterization of protein arginine methyltransferase of TgPRMT5 in *Toxoplasma gondii*. *Parasit Vectors* 12(1):221
 57. Dowden J, Pike RA, Parry RV, Hong W, Muhsen UA, Ward SG (2011) Small molecule inhibitors that discriminate between protein arginine N-methyltransferases PRMT1 and CARM1. *Org Biomol Chem* 9(22):7814–7821
 58. Yakubu RR, Silmon de Monerri NC, Nieves E, Kim K, Weiss LM (2017) Comparative monomethylarginine proteomics suggests that Protein Arginine Methyltransferase 1 (PRMT1)

- is a significant contributor to arginine monomethylation in *Toxoplasma gondii*. *Mol Cell Proteomics* 16(4):567–580
59. Cui L, Fan Q, Cui L, Miao J (2008) Histone lysine methyltransferases and demethylases in *Plasmodium falciparum*. *Int J Parasitol* 38(10):1083–1097
 60. Cui L, Miao J (2010) Chromatin-mediated epigenetic regulation in the malaria parasite *Plasmodium falciparum*. *Eukaryot Cell* 9(8):1138–1149
 61. Shi Y, Whetstone JR (2007) Dynamic regulation of histone lysine methylation by demethylases. *Mol Cell* 25(1):1–14
 62. Volz J, Carvalho TG, Ralph SA, Gilson P, Thompson J, Tonkin CJ, Langer C, Crabb BS, Cowman AF (2010) Potential epigenetic regulatory proteins localise to distinct nuclear sub-compartments in *Plasmodium falciparum*. *Int J Parasitol* 40(1):109–121
 63. Kaur I, Zeeshan M, Saini E, Kaushik A, Mohammed A, Gupta D, Malhotra P (2016) Widespread occurrence of lysine methylation in *Plasmodium falciparum* proteins at asexual blood stages. *Sci Rep* 6:35432
 64. Li H, Child MA, Bogoy M (2012) Proteases as regulators of pathogenesis: examples from the Apicomplexa. *Biochim Biophys Acta* 1824(1):177–185
 65. Withers-Martinez C, Jean L, Blackman MJ (2004) Subtilisin-like proteases of the malaria parasite. *Mol Microbiol* 53:55–63
 66. Blackman MJ, Fujioka H, Stafford WHL, Sajid M, Clough B, Fleck SL, Aikawa M, Grainger M, Hackett F (1998) A subtilisin-like protein in secretory organelles of *Plasmodium falciparum* merozoites. *J Biol Chem* 273:23398–23409
 67. Sajid M, Withers-Martinez C, Blackman MJ (2000) Maturation and specificity of *Plasmodium falciparum* subtilisin-like protease-1, a malaria merozoite subtilisin-like serine protease. *J Biol Chem* 275:631–641
 68. Harris PK, Yeoh S, Dluzewski AR, O'Donnell RA, Withers-Martinez C, Hackett F, Bannister LH, Mitchell GH, Blackman MJ (2005) Molecular identification of a malaria merozoite surface sheddase. *PLoS Pathog* 1:241–251
 69. Green JL, Hinds L, Grainger M, Knueffer E, Holder AA (2006) Plasmodium thrombospondin related apical merozoite protein (PTRAMP) is shed from the surface of merozoites by PfSUB2 upon invasion of erythrocytes. *Mol Biochem Parasitol* 150:114–117
 70. Lagal V, Binder EM, Huynh MH, Kafsack BF, Harris PK, Diez R, Chen D, Cole RN, Carruthers VB, Kim K (2010) *Toxoplasma gondii* protease TgSUB1 is required for cell surface processing of micronemal adhesive complexes and efficient adhesion of tachyzoites. *Cell Microbiol* 12:1792–1808
 71. Miller SA, Thathy V, Ajioka JW, Blackman MJ, Kim K (2003) TgSUB2 is a *Toxoplasma gondii* rhoptry organelle processing proteinase. *Mol Microbiol* 49:883–894
 72. Dowse TJ, Pascall JC, Brown KD, Soldati D (2005) Apicomplexan rhomboids have a potential role in microneme protein cleavage during host cell invasion. *Int J Parasitol* 35:747–756
 73. Brossier F, Jewett TJ, Sibley LD, Urban S (2005) A spatially localized rhomboid protease cleaves cell surface adhesins essential for invasion by *Toxoplasma*. *Proc Natl Acad Sci USA* 102:4146–4151
 74. Baker RP, Wijetilaka R, Urban S (2006) Two Plasmodium rhomboid proteases preferentially cleave different adhesins implicated in all invasive stages of malaria. *PLoS Pathog* 2:922–932
 75. Francis SE, Banerjee R, Goldberg DE (1997) Biosynthesis and maturation of the malaria aspartic hemoglobins plasmepsins I and II. *J Biol Chem* 272:14961–14968
 76. Liu J, Gluzman IY, Drew ME, Goldberg DE (2005) The role of *Plasmodium falciparum* food vacuole plasmepsins. *J Biol Chem* 280:1432–1437
 77. Luker KE, Francis SE, Gluzman IY, Goldberg DE (1996) Kinetic analysis of plasmepsins I and II, aspartic proteases of the *Plasmodium falciparum* digestive vacuole. *Mol Biochem Parasitol* 79:71–78
 78. Singh N, Sijwali PS, Pandey KC, Rosenthal PJ (2006) *Plasmodium falciparum*: biochemical characterization of the cysteine protease falcipain-2. *Exp Parasitol* 112:187–192

Part IV

Infection Metabolomics: Metabolomics Applications for Human Pathogens



An Introduction to Computational Pipelines for Analyzing Untargeted Metabolomics Data for Leishmaniasis **21**

Anita Verma, Arunangshu Das, and Chinmay K. Mukhopadhyay

Abstract

Metabolomics helps us to understand the metabolism of biological systems and also metabolic interaction between host and parasites. This chapter describes the methodology of metabolite extraction and derivatization from the biological samples with reference to leishmaniasis. Here, we also provide stepwise installation of various applications on R platform such as MetaboAnalyst, CDK-R, and some other supporting package for processing and analyzing of untargeted metabolomics data. These packages are used to perform mass spectral peak binning, identification of metabolites/pathways and finally development of pipeline to identify, design, and perform chemometric analysis that may be helpful in designing drug against potential targets.

Keywords

Metabolomics · Leishmaniasis · R-programming

A. Verma (✉)

School of Life Science, Jawaharlal Nehru University, New Delhi, India

A. Das

CSIR-Institute of Genomics and Integrative Biology, New Delhi, India

C. K. Mukhopadhyay

Special Center for Molecular Medicine, Jawaharlal Nehru University, New Delhi, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_21

21.1 Introduction

Infectious parasitic diseases are very complex. Protozoan parasite-derived diseases like leishmaniasis, malaria, Chagas disease, and sleeping sickness are widely known and studied. According to WHO 2020 report, *Leishmania* continues to spread, with a total of 97 endemic countries or territories, of which four were added in 2017–2018 [1]. An estimated 30,000 new cases of visceral leishmaniasis and more than 1 million new cases of cutaneous leishmaniasis occur annually. Despite its widespread infection in tropical countries, it remains as one of the most neglected diseases. Visceral leishmaniasis caused by *Leishmania donovani* (LD) or *Leishmania infantum* is the most severe form of the disease. Lack of treatment causes 95% fatality [2]. *Leishmania* is an obligate parasite. In the gut lumen of sandfly vector, it multiplies and transforms into motile and infectious metacyclic promastigote, whereas in the vertebrate host, the promastigote form differentiates into an obligatory intracellular amastigote within macrophages [3].

Host-parasite interaction plays a very important role for survival of parasites within the host cell. A significant metabolic interaction between parasites and host is required to divert host's nutrients toward the growth and multiplication of the parasites while the host struggles to maintain their homeostasis and cope with waste products, toxins, and associated tissue damage [4]. Over the last decade, our understanding of parasitic diseases has been immensely benefitted by the study of omics, and metabolomics is one of them.

Despite the fact that metabolomics is an emerging and highly powerful technique to understand the host-parasite interactions [5], it also poses a significant challenge for application and to derive appropriate implications. It requires extensive data curation and careful analysis to obtain logical and significant outcomes. These aspects render metabolomics highly dependent on process analytics and statistics. In this chapter, we will briefly review the process pipelines available for metabolomics data analysis. We will discuss the useful codes and their meanings, which should help in understanding these pipelines, and perform computation bypassing the extensive mathematics and by learning minimal programming. We have focused on pipelines which are based on R because they are simple, freely available, and backed by assistance of a dynamic community involved in making the platform more accessible and user-friendly.

21.1.1 Recent Advances in Understanding Parasitic Disease Using Metabolomics with Reference to *Leishmania*

There has been significant increase in literature resources regarding metabolomics in the field of parasitosis. Leishmaniasis is one of the most important neglected tropical diseases around the world and has also been studied on the metabolomics platform. The burden of leishmaniasis in the world health scenario is significant especially in the low-income countries. Systematic studies over the years have clearly shown the changing nature of the parasite and its evolving capacity of developing drug

resistance. The treatment of *Leishmania* still heavily relies on chemotherapeutic agents; the traditional frontline drugs like pentavalent antimonials suffer from both high host toxicity and progressive drug resistance of the parasite. The effort toward understanding the metabolomic network of *Leishmania* started with the reconstruction of the metabolic network of *L. major* that accounted for 560 genes, 1112 reactions, 1101 metabolites, and eight unique subcellular localizations [6]. Using $^1\text{H-NMR}$ -based metabolomics, it was possible to identify a list of metabolites like citraconic acid, isopropylmalic acid, L-leucine, ornithine, and caprylic acid. These metabolites were altered differentially during the logarithmic and stationary growth phase of promastigote [7]. A comparative study of the metabolome profile of promastigotes of *L. donovani*, *L. major*, and *L. mexicana* was performed by LC-MS technique, and the identity of 64 metabolites differing in threefold level or more between the cell extracts of species were confirmed [8]. Significant deregulation in metabolic pathways were observed for amino acids biosynthesis. Trp, Asp, Arg, and Pro levels were found to be altered in the global metabolome pool of the studied *Leishmania* spp. [8]. Primarily, targeted metabolomics have been performed on whole leishmania cell lysate or host tissues. The targeted metabolomics primarily aims at understanding the corresponding difference in metabolome profiles at various stages of growth or the markers of transition between different phases in the life cycle.

Untargeted metabolomics performed on *Leishmania* or other parasites has more often been involved in figuring out alterations that are not fully understood or predicted. Thus, untargeted metabolomics requires intensive analysis of the data obtained to map significant changes. Most of the available untargeted metabolomics data for *Leishmania* have been obtained after application of chemotherapeutic treatments against suitable controls.

Vincent et al. reported around 80 metabolites that have been significantly altered through performing untargeted metabolomics of miltefosine-treated *Leishmania infantum* JPCM5 promastigote cell line [9]. Pountain and Barrett have recently reported phenotypic variation in amphotericin B-resistant *Leishmania* parasites using untargeted metabolomics [10]. The raw data obtained from *Leishmania*-untargeted metabolomics have mostly been analyzed on respective instrument-based software platforms. There are also a few reports where the data have been analyzed on instrument-independent platforms like XCMS. The primary parameters that mostly concern investigators on XCMS platform are signal to noise ratio (usually set to 2), bandwidth (usually set to 2), and minimum fraction of samples necessary in one group to be a valid group (usually set at 0.25), while other parameters are set as default. Multivariate analysis used with PCA and PLS-DA are used for studying the difference between the groups. Metabolites are identified using Human Metabolome Database, Kyoto Encyclopedia of Genes and Genome (KEGG), and Metlin databases using $[\text{M} + \text{H}]^+$, $[\text{M} + 2\text{H}]^{2+}$, and $[\text{M} + \text{Na}]^+$ as possible adducts and 5 ppm as maximum error.

It is important to understand that mass spectrometry data for untargeted metabolomics heavily depend on metadata for successful interpretation. This gives rise to the possibility of variable approach to analyze the same data using different

mathematical approaches for obtaining results. This chapter is dedicated to unlock the basic themes of some of the tools developed for analysis of untargeted metabolomics but has been rarely used for understanding biology of *Leishmania* or leishmaniasis. This should broaden the scope of understanding of host-parasite interaction in leishmaniasis and will also introduce the very basics of drug development using the result of untargeted metabolomics [11].

21.1.2 Metabolomics Pipeline

Metabolomics pipeline is simple but there exists wide scope of approaches and strategies. It is mainly performed with two clear objectives:

- To quantify targeted metabolites and understand the biological significance of altered metabolites
- To perform total metabolome scan and obtain qualitative picture of metabolome perturbation to narrow down target region to perform the previous objective

The first objective is usually performed when there is prior information about the changes or probable changes available with the researcher. Absolute quantification provides the scope to construct metabolic models that help researchers design experiments without the need of performing metabolomics repeatedly. Metabolic models are also helpful in predicting the effect of targeted changes like deletion of pathway enzymes and targeting cells' transcriptional control devices. Such highly accurate models are already available for simple organisms like *E. coli* and yeast, but due to the inherent complexity of cellular organization and extensive differentiation, such models are not still widely available for higher organisms like human. Research is already progressing to build such models for individual tissues, but that not only requires sophisticated instrumentation and automation but also needs to be complemented with significant computational capabilities.

Untargeted metabolomics bridges the vital gap between absolute quantification of downstream metabolites and upstream genomics data. The relation between transcription, translation, and metabolism is highly complex. Metabolites are the terminal product of the above mechanisms, and their levels are real pictures of the existing state of cells. They are not only involved in catabolic and anabolic processes, pH regulation, and redox homeostasis, but they are also involved in less-understood functions like protein folding, cellular proteostasis capacity, and DNA-RNA-protein interactions. Thus, untargeted metabolomics can uncover biological information to complement other omics highlights. We will analyze each of these steps to provide a clear idea for implementing metabolomics in disease research (Fig. 21.1).

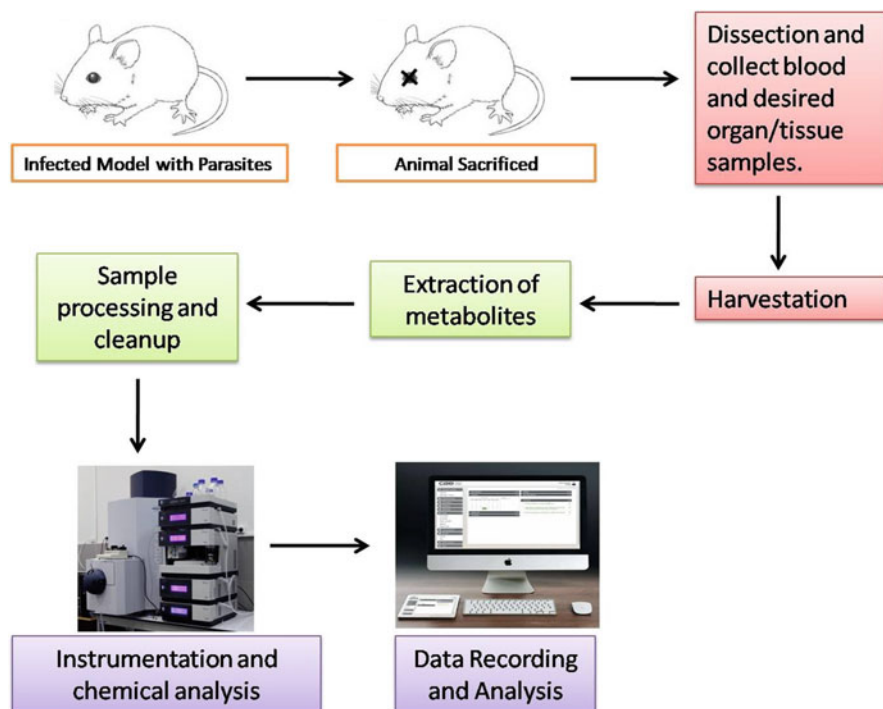


Fig. 21.1 Schematic diagram of summary of metabolomics analysis pipeline

21.2 Extraction, Derivatization, and Processing of Metabolome from Host at Different Postinfection Phases

21.2.1 Extraction of Metabolites and Derivatization

Metabolites produced by metabolic pathway, affected by diseases, may serve as biomarkers for respective infectious disease. Carbohydrates, amino acids, nucleic acids, lipids, and hormone levels are usually measured for understanding particular parasitic diseases. LC-MS, GC-MS, and NMR are used for the identification of metabolite in qualitative and quantitative levels [12]. For extraction of metabolites from various biological targets, a repository of protocols is available online at the following link: <https://www.metabolomicsworkbench.org/databases/metabolitedatabase.php>. The following is the summary of metabolite extraction protocol from parasites grown in vitro and from host body fluids (such as blood, serum, and urine) and tissues.

Box 21.1 Metabolite extraction protocol

In vitro-grown parasite culture	Host plasma/serum sample
<ol style="list-style-type: none"> 1. Collect parasites containing the respective media into the centrifuge tube. 2. Keep the parasite at 0 °C in pre-chilled ethanol. 3. Centrifuge to separate media from parasites on 5000 rpm at 0 °C for 10 min. 4. Discard supernatant, and suspend the pellet into pre-chilled PBS. 5. Remove remaining media by washing of pellet three times with 1 ml pre-chilled PBS, and centrifuge at 5000 rpm at 0 °C for 5 min (after every centrifugation, remove supernatant thoroughly with pipette tips). 6. Number of parasites counted by using hemocytometer. 7. Take 4×10^7 cells, and add 500 μl of pre-chilled extraction solution (acetonitrile/methanol/MiliQ, 2:2:1) keeping always at 0 °C. 8. Cell destruction and metabolite extraction on 1400 rpm at 4 °C for 1 h by using ThermoMixer. 9. Centrifuge the sample for 15 min at 14,000 rpm at 4 °C. 10. Collect supernatant into fresh new glass vial. 11. Deoxygenate the extracted sample with a gentle stream of nitrogen gas for 1 min prior to vial closure, and store at –80 °C until further use. 	<ol style="list-style-type: none"> 1. Collect blood from the host, and separate the plasma/serum from blood. 2. Vortex the plasma/serum samples for 10 s. 3. Aliquote 40 μl and add 160 μl of pre-chilled extraction solution (acetonitrile/methanol/MiliQ, 2:2:1) maintaining 0 °C always. 4. Vortex for 30 s, and shake for 5 min at 4 °C using ThermoMixer. 5. Keep at –20 °C for 20 min. 6. Centrifuge for 15 min at 14,000 rpm at 4 °C. 7. Transfer 200 μl of supernatant in the glass vial. 8. Deoxygenate the extract with a gentle stream of nitrogen gas for 1 min prior to vial closure, and store at –80 °C until further use.

Box 21.2 Derivatization protocol for GCMS/MS

1. Prepare 40 mg/ml methoxyamine HCl solution in pyridine.
2. Evaporate the extracted sample described in the Box 21.1 using gentle stream of nitrogen gas until complete dryness.
3. Add 40 μ l of methoxyamine HCl solution to the dried sample.
4. Incubate at 37 °C for 90 min.
5. Add 80 μ l of BSTFA reagent. Vortex for 10 s.
6. Incubate at 70 °C for 60 min.
7. Transfer the content to fresh glass vial after cooling at room temperature, and store at –80 °C until further use.

For LC-MS instruments, extracted sample can be directly injected on column described in the Box 21.1.

For GC-MS instruments, derivatization step is required that has been described in the Box 21.2.

21.2.2 Computational Pipeline for Metabolomics Data Analysis

The workflow has been described for the installation of R version 4.0.2 with reference to MetaboAnalyst, which is supported by R version 4.0.2 and above. The latest version of R software (R version 4.0.2) is freely available for Linux, Windows, and Mac OS X on their websites (<https://cran.r-project.org/>). It is important to note that R version 4.0.2 [13] does not come with “R developer tools” (`devtools`); hence, installation of older packages is not possible in R version 4.0.2 using “`devtools`” [14]. Many of the computational packages for analysis of metabolomics data, e.g., MetaboAnalyst, are under evolutionary status on R version 4.0.2. Users face multiple problems due to version incompatibility of R packages which requires IT consultation and extensive collaboration. Online assistance like the “Bioconductor program” provides vital inputs and can be consulted as often as required. To enlighten the user, here, we describe step-by-step installation procedure for the tools and MetaboAnalyst 3.0 (<https://www.metaboanalyst.ca/docs/RTutorial.xhtml>), which is freely available in GitHub. An online version of this platform is available, but running this on personal systems provides more freedom to user for data jiggling.

21.2.2.1 Rtools Installation

- R 4.0.0 and above versions have been released after April 2020; R for windows comes with a bundle of tool chain named “`rtools40`”. This version of Tools upgrades the `mingw_w64 gcc` tool chains to version 8.3.0 and introduces a new build system on `msys2`.
- To use `Rtools40`, download the installer from CRAN using the links “`rtools40-x86_64.exe`” and “`rtools40-i686.exe`” for 64 bits. User should ensure using only RStudio version 1.2.5042 or above to work with `rtool40`.
- Install `Rtools` for compiling R packages. It is required to put the location of `Rtools` and make utilities (`bash`, `make`, etc.) on the `PATH`. An easy option is to direct all R-related installation in the same directory. Type commands on the `Rtools` terminal as mentioned in Fig. 21.2.
- Now, start R programming R.0.2. To look at the `PATH` of `Rtools`, type the command as shown in Fig. 21.3:

```
>sys.which ("make")
```

The output should look like below:

```
>C:\crrtools40\cruser\crbin\crmakes.exe
```

```

Arunangshu@LAPTOP-DH3NL6JU MSYS ~
$ PATH="{RTOOLS40_HOME}\usr\bin;${PATH}"

Arunangshu@LAPTOP-DH3NL6JU MSYS ~
$ git clone https://github.com/xia-lab/MetaboAnalystR.git
Cloning into 'MetaboAnalystR'...
remote: Enumerating objects: 146, done.
remote: Counting objects: 100% (146/146), done.
remote: Compressing objects: 100% (87/87), done.
remote: Total 3885 (delta 84), reused 101 (delta 59), pack-reused 3739
Receiving objects: 100% (3885/3885), 73.65 MiB | 1.96 MiB/s, done.
Resolving deltas: 100% (2690/2690), done.

Arunangshu@LAPTOP-DH3NL6JU MSYS ~
$ CMD build MetaboAnalystR
Microsoft Windows [Version 10.0.18362.1016]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\Arunangshu>CMD INSTALL MetaboAnalystR_3.0.0.tar.gz
CMD INSTALL MetaboAnalystR_3.0.0.tar.gz
Microsoft Windows [Version 10.0.18362.1016]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\Arunangshu>

```

Fig. 21.2 Process of command PATH of Rtools

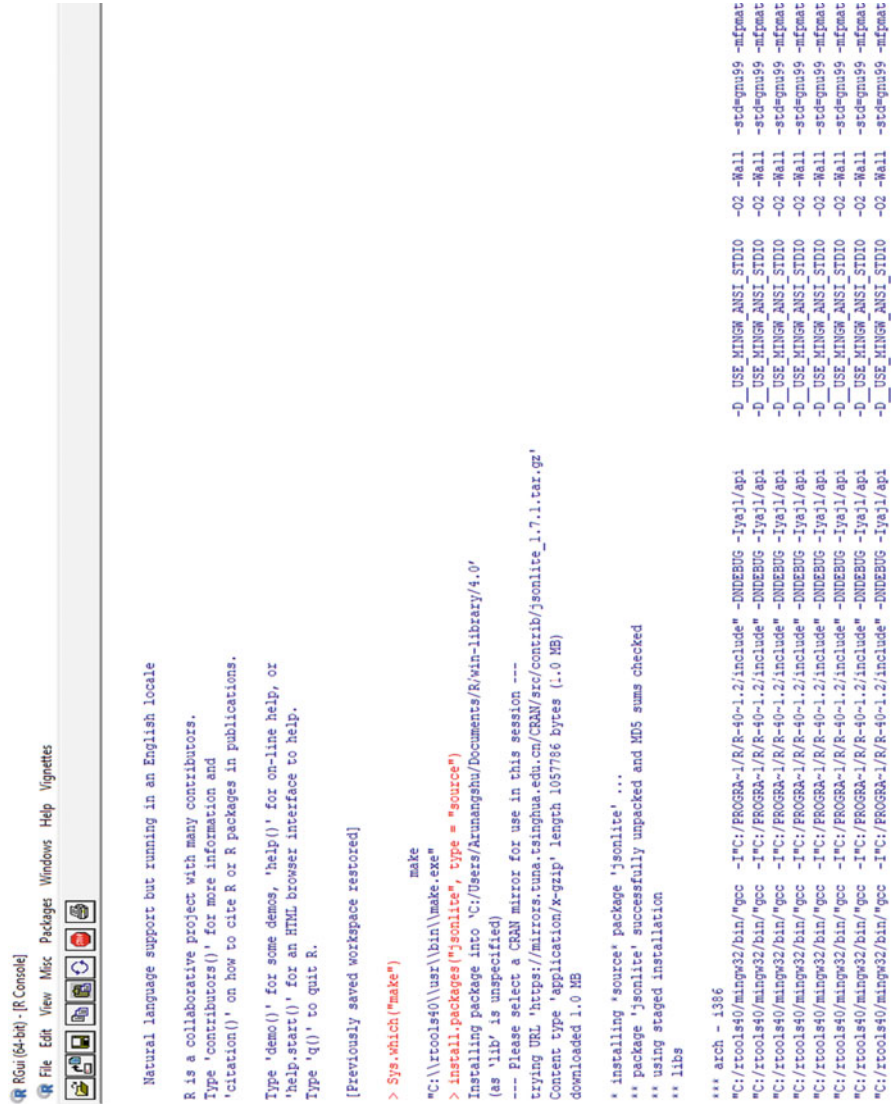


Fig. 21.3 Verification of command PATH of Rtools in R console

The successful batch of Tools was verified by installing “jsonlite” [15]. It can be seen that jsonlite has been successfully downloaded and installed (Fig. 21.3). The R package of MetaboAnalyst was directly cloned from source (Fig. 21.2).

21.2.2.2 Installation of MetaboAnalyst

(i) The package MetaboAnalyst 3.0 is available as an R package. It can be installed by three different ways:

- As previously described through "devtools"
- Cloning the GitHub
- Manually downloading the .tar.gz file

As we have discussed that for R 4.0.2 and onwards devtools are not available, we shall here proceed through cloning GitHub directly through mirrors online. In addition, we have attempted to install MetaboAnalyst 3.0 using the manually downloaded .tar.gz file, but it was not successful. In order to clone GitHub, open the "bash" window of Rtools, and write the following command for git clone:

```
>http://github.com/xia-lab/MetaboAnalystR.git
```

Press "ENTER."

The immediate output should say cloning into "MetaboAnalystR"; this should be followed by three remote processes: enumeration, counting, and compressing. After the output shows done, type the command:

```
>CMD build MetaboAnalystR
```

After the code has successfully been executed the compiler will return the default title of Windows, e.g., "Microsoft Windows (version 10.0.18362.1016)." The user will be automatically directed to the target folder. This is performed using the command:

```
>CMD INSTALL MetaboAnalystR_3.0.0.tar.gz
```

When this command is executed, now, MetaboAnalyst is in line to be executed as an R package in R console.

Now, MetaboAnalyst has to be installed on the R platform. This can be again done by two ways:

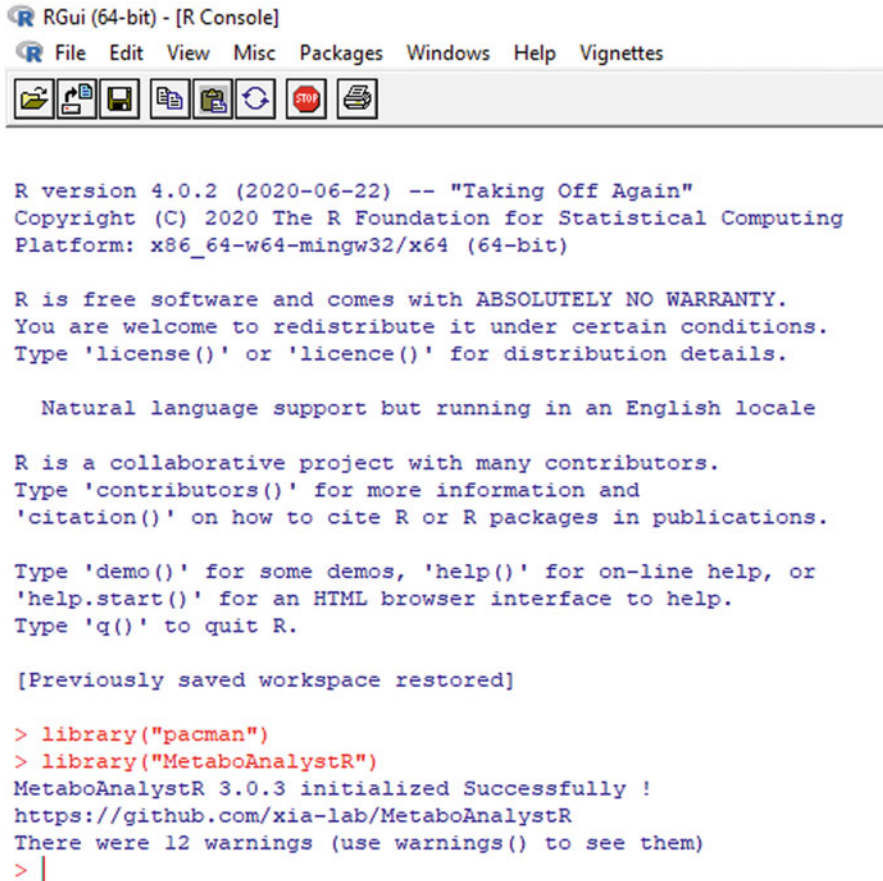
- By installing "metanr_packages" [16]
- Installing "pacman" [17]

Since we are in R platform 4.0.0 and above, it is recommended to use "pacman". Run the R command:

```
>install.package ("pacman")
```

To use the installed package, run the command:

```
>library ("pacman")
```



```

R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> library("pacman")
> library("MetaboAnalystR")
MetaboAnalystR 3.0.3 initialized Successfully !
https://github.com/xia-lab/MetaboAnalystR
There were 12 warnings (use warnings() to see them)
> |

```

Fig. 21.4 Successful installation of MetaboAnalyst displayed in R console

The “pacman” packages contain important functions like the following:

“multest” – Nonparametric boot strap and permutation resampling-based multiple testing procedures (including empirical Bayes methods) for controlling family-wise error rate (FWER), generalized family-wise error rate (gFWER), tail probability of the proportion of false positive (TPFP), and false discovery rate (FDR) [18].

“genefilter” – Methods for filtering genes from high-throughput experiments [19].

“globaltest” – Testing groups of covariates/features for association with a response variable, with applications to gene set testing [20].

After installation during subsequent uses, just enter the command:

```
>library ("pacman")
```

any time for using on R. Now, directly load MetaboAnalyst on R by using the command:

```
>library ("MetaboAnalystR")
```

Now, the R console should show successful installation and initialization of MetaboAnalyst 3.0.3. This concludes the installation of MetaboAnalyst on your PC (Fig. 21.4).

The system may show multiple warnings which are good to read and understand but can be set aside unless some serious errors do not pop up.

21.3 Statistical Analysis of Metabolome Data

The main challenge of the metabolomics study is data analysis to recognize and identify all metabolites and their significant alterations to confirm biomarkers and interpret relevant biological significance.

There are many freely online software available for the statistical analysis, data analysis, and graphics like MATLAB [21], XCMS [22], MetaboAnalyst [23], Haystack [24], MZmine2 [25] and Metabolomics Ion-Based Data Extraction Algorithm (MET-IDEA) [26], and finally R packages. The MetaboAnalyst online website has a size restriction of up to 50 M; the R package will be of great use to users for both direct processing and batch processing of larger datasets.

21.3.1 Data Conversion

In the case of LC-MS analysis, data generated from each chromatogram are arranged in datasets containing information of mass-to-charge ratio (m/z), retention times, and intensities. When working in R environment, LC-MS data are usually imported by means of the mzR package available at Bioconductor (<https://www.bioconductor.org/packages/release/bioc/html/mzR.html>). mzR provides a unified interface for most of the open data formats described above such as mzXML, mzML, mzData, and netCDF [27]. Afterward, *peaks* function can be used to extract all MS spectral data into a matrix to be further analyzed (i.e., XCMS and MSnbase [28]). The first step in data processing of targeted and untargeted metabolomics is to convert the raw data into a numerical format that can be used for downstream statistical analysis.

21.3.2 Preprocessing

For LC-MS data, useful preprocessing steps include data filtering, feature detection, alignment, and normalization [29].

21.3.2.1 Normalization

In biological samples, the concentrations of metabolites depend on a large number of factors, many of which are not a concern of metabolomics analysis. Rather than directly going for XCMS or MetaboAnalyst beginners can start handling small datasets using many freely available simple pipelines from the R community. For example, the use of MALDIquant package for analysis of MALDI-TOF and 2D mass spectrometry data [30] is highly recommended in the beginning. The package includes all essential steps of data processing pipelines such as raw data import, baseline removal, and peak detection. To start with, we have to install the package. The package is compatible with any version of R above 3.2.0 and can be simply installed by the command:

```
>install.packages("MALDIquant")
```

There are multiple approaches to normalize datasets in metabolomics including total signal normalization and vector normalization. The total ion current (TIC; sum of all intensities under curve) is computed on raw data, baseline corrected data, smooth data time scale, and m/z scale. However, on m/z scale, computation can fail due to the fact that total area under the curve can be very large. Thus, normalized values are required to be multiplied by large constant to put intensities on a scale. Alternative methods for normalization include probabilistic quotient normalization (PQN) [31] or median. For example, the standard R code for PQN normalization using MALDIquant [32] is:

```
>spectra<-calibrateIntensity(objectname, method=c("TIC", "PQN",  
"median"),range=c(3000, 5000))
```

or

```
>spectra<-calibrateIntensity(objectname, method="TIC", range=c  
(3000, 5000))
```

If range is provided, TIC is only calculated for the specified mass range. The value returned to spectra in the upper code is an object with calibrated intensities.

21.3.2.2 Baseline Correction

The baseline correction is performed to get rid of the noises in the signal. Maintenance of instrument is essential along with clean sample preparations. The general practice of determining and correcting baseline is to keep the signal to noise ratio between 2 and 5. The algorithm is based on SNIP described by Ryan and his colleagues [33]. The algorithm is based on the following equation:

$$y_i(k) = \min y_i, \frac{(y_i - k + y_i + k)}{2}$$

where $y_i(k)$ is mean channel value, y_i is channel value of spectra, and k denotes spectral width.

The algorithm gives acceptable spectra after only 24 iterations; however, for mass spectrometry data, 100 iterations are generally performed. Using MALDIquant package, the following lines of codes can be executed for baseline correction:

```
>b1<-estimateBaseline(objectname, method="SNIP", iterations=n)
plot (b1,col="red", lwd=2)
```

Multiple iterations should be performed for desired baseline, e.g., $n = 75$, 100, 150, etc.

To remove baseline:

```
>b1<-removeBaseline(objectname, method="SNIP", iterations=100)
plot (b1,col="red", lwd=2)
```

21.3.2.3 Smoothing

The spectral intensity can be smoothened using Gaussian-Lowess and Savitzky-Golay [34]. The filter coefficient in Savitzky-Golay is derived by performing unweighted linear least square fit using a polynomial of higher degree which preserves signal feature like resolution between two peaks and height of the peak.

The smoothing can be performed by using MALDIquant:

```
>b1<-smoothIntensity (objectname, method="SavitzkyGolay",
halfWindowSize=10)
```

21.3.2.4 Peak Detection

Prior to peak detection, it is important to estimate the noise in the spectra. Using the following algorithm, one can determine and plot the noise and the signal:

```
> noise <- estimateNoise (objectname)
> plot(objectname, xlim=c(2000, 5000), ylim=c(0, 0.002))
> lines(noise, col="select colour")
> lines(noise[,1], noise[, 2]*2, col="select colour")
```

Here, the X -axis represents the m/z values that have been set between 2000 and 5000. The normalized signal intensity has been similarly set between 0 and 0.002. The user is free to select any desired range. Once the signal and noise ratio is decided, peaks can be detected using MALDIquant [35]:

```
> selectpeaks<-detectPeaks(avgSpectra, method="MAD",
halfWindowSize=20, SNR=2)
> plot(avgSpectra, xlim=c(2000, 5000), ylim=c(0, 0.002))
> points(peaks, col="select colour", pch=4)
```

In MetaboAnalyst, after successful installation of the package followed by library of the package, data can be inspected using the command:

```
>"PerformDataInspect(directoryname, res=50)"
```

This will provide the RT range and the m/z range. The command

```
>"PerformParamsOptimization()"
```

accepts data file name, initialization, and number of CPU cores to be assigned for the job as default arguments. The single command determines signal to noise ratio, smoothing using Gaussian-Lowess, and performing mean centering by subtracting the mean value from each peak for all the same peak across the sample. The mean value of each peak is determined by averaging values across the column and then subtracting the mean from each peak belonging to different samples across the row.

21.4 Identification of Target Metabolic Pathway and Biomarker

There are two main approaches to decipher metabolic alteration. One approach is to identify metabolites that are altered in test condition compared to control or in untargeted metabolomics by finding significant features that are altered across the sample space and then constructing logical trees to reach conclusion about pathways using biological experience. The former approach has been quite successful and has led to numerous important outcomes. However, the process has certain limitations which include user bias towards conclusions. Also, for very large datasets, such comparison becomes nearly impossible. Another approach is to skip identification of metabolites and directly look for variations in clusters of daughter ions and directly infer pathways. However, this process indirectly depends on initial metabolite identification to build the pipeline for meta-analysis of ions to pathways directly. Here, a brief discussion will be presented about both these practices. There is no hard rule to prefer one approach over the other. A lot of scientific literature is available about the merits and demerits of these approaches individually; however, the overlap of outcome of these two approaches depends also on sample quality, sample preparations, instrument maintenance, and the fine tuning of hardware.

21.4.1 Identification of Metabolite with XCMS

Metabolites are identified in XCMS software which is freely available online through METLIN database (https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage) and KEGG compound IDs (<https://www.genome.jp/kegg/compound/>). METLIN contains MS/MS data for more than 12,000 metabolites and 61,000 MS/MS spectra. Matches with MS/MS data are scored using a cosine similarity metric (0–1) to the closest collision energy in the database. Spectral mirror plots are created for confirmation of positive matches. Modern mass spectrometry workflow can also be a supervised analysis where the user's knowledge is applied to select target. A biological sample is initially separated on column, and the compound enters a single quadrupole mass spectrometer or in triple quadrupole mass spectrometer operating in simple ion scanning of the mother ion. From the peak area, a relative quantification is performed, and the METLIN database is searched; next, the mother ion is fragmented, and the relative intensities of daughter ion are determined. Based on the m/z values of the daughter ions, the mother ion and hence the compound can be identified. A user-supervised quantification of the compound can be possible by splitting the mother ion at different energies and looking for fragments. Each fragment is then selected, and based on trial and error method, the user-defined fragment is repeatedly scanned at the last quadrupole to achieve quantification. These methods have been used for both identification and quantification of the compounds. In XCMS, once the submitted job is completed, the most important feature is the result table which can be used by the user to perform various analysis. The data table contains initially a large number of entries that can be filtered using p value, fold change, and max intensity filters as shown in Fig. 21.5.

Appropriate choice of filters like p value set to <0.005 , fold change set to >10 , and maximum intensity set to above 10,000 filters out a lot of unnecessary entries and highlights more significant ones. A note of caution is that filter parameters depend on what the observer is looking at; it is not a universal set parameter for every application.

21.4.2 From Mass Spectral Peaks to Pathways Bypassing Metabolite Identification

Untargeted metabolomics requires identification of metabolites which is a major challenge. Determination of metabolite identity only based on the m/z values is an overwhelming approximation. The idea of bypassing metabolite identity is not to ignore metabolites, but to bypass the steps of identification to decipher pathway and metabolic network. This ultimately requires the unification of metabolite identity and pathway identity simultaneously rather than piggybacking pathway analysis on identification. This approach is based on the hypothesis that if a biologically significant change is manifested in cellular metabolome, there should be local enrichment of metabolites in the pathway or network, while the outliers will be distantly related. Thus, if the aromatic amino acid biosynthesis pathway is

The screenshot displays the XCMS software interface. At the top, there is a navigation bar with options like Home, MEM, Databases, Create Job, View Results, XCMS Public, XCMS Institute, Stored Datasets, Account, Toolbox, Help, and Logout. Below this, a job summary for 'Job#1205179 : P_2019-01-22_22:34' is shown. The main area contains a table with columns: field, value, up/down, named, rt/med, ms/act, dataset1_mean, dataset2_mean, isotopes, adducts, peakgroup, and usernotes. The table is filtered to show records where 'value' is greater than 0.01 and 'fold' is greater than or equal to 10. A search filter is visible on the left, showing 'value' set to 'less' with a value of '0.01' and 'fold' set to 'greater or equal' with a value of '10'. The table shows 174 records, with the first few rows having values like 96.1, 32.7, 40, 43, 45, 82, 66, 103, 128, 140, and 174. The interface also includes a footer with contact information for the Toxigen Research Institute.

Fig. 21.5 Setting up filters to remove and resize the table into significant and manageable form in XCMS

upregulated, an enrichment analysis can be performed using the possible peaks that manifest upregulation of prephanate, shikimate, and erythrose 4-phosphate rather than perturbations in distantly related sedoheptulose pathway and vice versa [36]. This function is thus performed by the “mummichog” algorithm (<https://pypi.org/project/mummichog/>) (Fig. 21.6). However, a certain limitation is that mummichog can search for all modules that can be built on user input data and compute their activity scores on a predefined reference metabolic network model. The basic test for pathway enrichment is Fisher’s exact test (FET), which is widely used in transcriptomic analysis [37]. The information about the pathway can be enumerated from the list of metabolites that is mapped with the list of m/z values. For every metabolite, this mapping is performed many times until significant enrichment is obtained and pathways are deduced by comparing the p values from the real data and the p values for the permutation mapping. Briefly, let’s consider a list of M metabolites. Now, if there exists a pathway with attribute A , first, m number of metabolites may be determined out of M that has been annotated with attribute A . Next, using Fisher’s exact test, the probability $p + (A)$ of having at least m such genes with attribute A is calculated if the null hypothesis is true. If $p + (A)$ value is sufficiently small, then one can consider m number of metabolites is statistically significant to pathway with A attributes. All the significant metabolites then can be used to draw a functional network (Fig. 21.7). Mummichog can be performed with

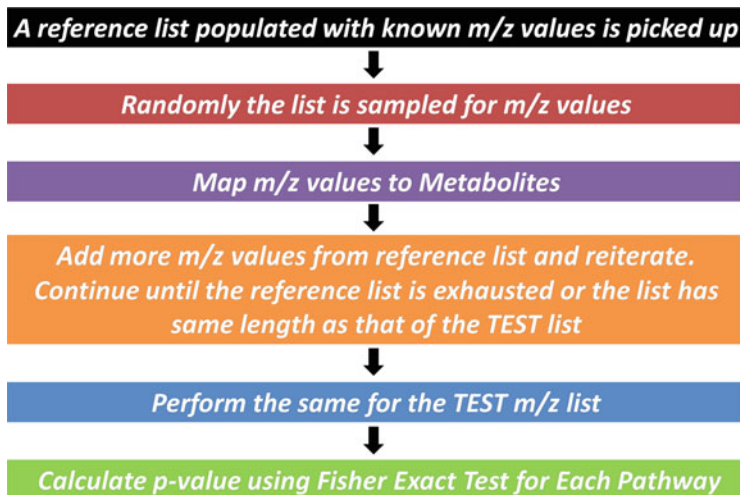


Fig. 21.6 Flow chart of mummichog logic

datasets on MetaboAnalyst Platform where fold change is considered as an attribute to construct the pathways that are significantly altered in test as compared to control.

21.5 Computer-Assisted Drug Design and Biochemical Validation

Currently, the development of new effective drugs is hindered not only by development costs, drug efficacy, and drug safety but also by the rapid occurrence of drug resistance/tolerance. Here, we discuss how metabolomics may help in identifying key metabolites in clinical conditions and subsequently led to development of possible new drug targets and drugs with an example provided below.

Atherosclerosis is the resulting clinical condition attributed to SNPs and high levels of cholesterol. In 2011, Wang et al. performed an untargeted metabolic profiling of rat plasma that has developed plaques. Compelling evidences were found between TMAO levels and developing atherosclerosis condition. TMAO is an oxidized by-product of trimethylamine (TMA) which is a product from metabolic breakdown of creatinine, betaine, choline, etc. by microorganism. Subsequent studies have also established strong correlation between atherosclerosis and TMAO in human. The identification of high levels of TMAO in plasma having a strong correlation between atherosclerosis has led to the identification of novel enzyme targets, i.e., the liver flavin monooxygenase III and bacterial TMA-lyase. Using computer-assisted docking, 3,3-dimethyl butanol was identified in olive oil which was able to inhibit TMA (choline)-lyase [38].

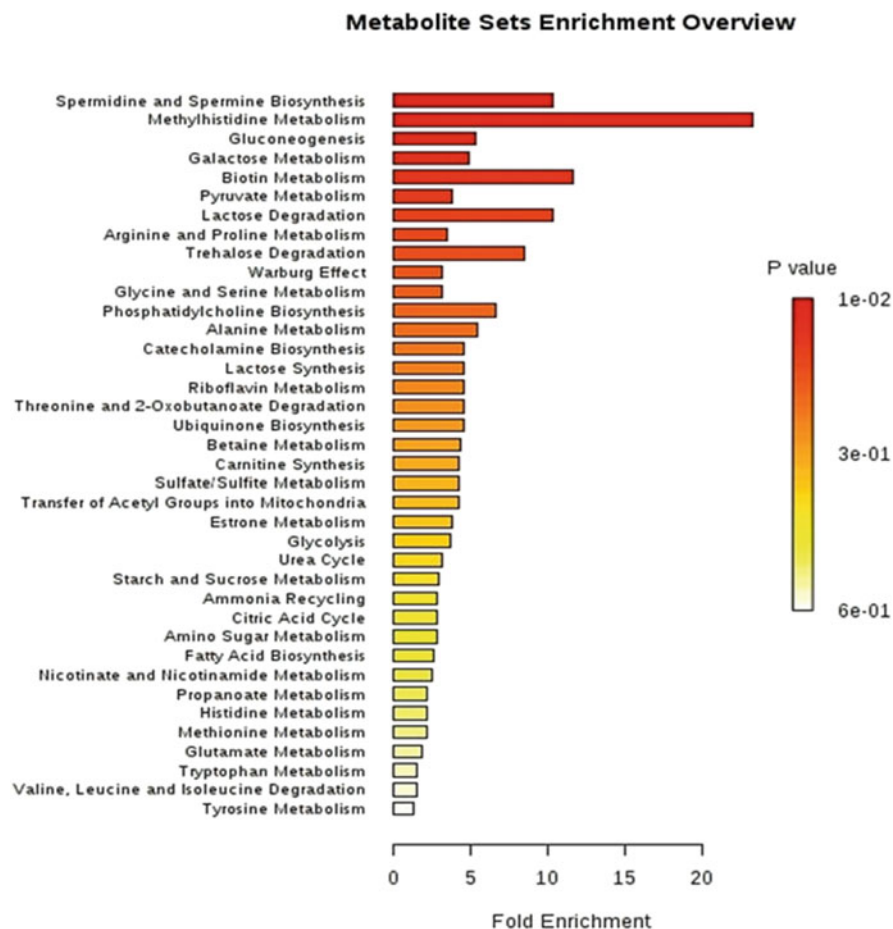


Fig. 21.7 A fold enrichment analysis performed using mummichog for pathway upregulation in test against control using MetaboAnalyst

Many of the best-selling drugs are enzyme inhibitors or antimetabolites. Thus, metabolomics clearly shows cost-effective time-saving direction toward identifying targets followed by design of drugs.

21.5.1 Selecting Target-Specific Drug Candidates

The process of selecting potential drug molecule is intensive. The search starts with the thousands of potential molecules which go through several phases of trials, and only a few entries make it to the final stage. This process is not only costly and time-consuming but often leads to deselection of potential candidates due to high

biological variability. Thus, a computer-assisted platform has been developed to facilitate the whole process. Here, we will discuss about R based platforms that utilize chemometrics and similarity index to narrow down potential target molecules with greater confidence and precision.

21.5.1.1 Molecule Exploration with "rCDK" Package on R

The name `rCDK` arises from Chemistry Development Kit (CDK), a collection containing free java libraries for wide variety of "Cheminformatics" functionality. It is used in R programming platform to provide direct access to CDK. The CDK and the R platform are combined in the package called "rCDK" [39]. However, `rCDK` package requires several other associated packages for execution of the program. To install a package in R, continue with the following command:

```
> install.packages("rCDK")
```

The associated packages require R "Chemometrics", "ChemmineR", "Cluster", "rgl", "vegan", "factoextra", "fingerprint", "fmcsR", "NbClust", "ggplot2", "iqspr", and "gridextra".

21.5.1.2 Brief Introduction of Some of the Packages

- "Chemometrics"—this deals with multicollinearity in data having a large number of variables and a very limited observation. Chemometrics rely on principle component analysis to transfer the original variables to a small set of principle components to maximize variance of the data. The package contains several methods of regression which are suitable for chemical data [40]. The package contains OLS, principle component regression, and partial least square (PLS) option. The package is a supervised training algorithm which classifies objects based on previous training session. Chemometrics on "rCDK" utilize simple methods like *k nearest neighbors* and *classification trees* to highly complex artificial neural networks [41].
- "ChemmineR"—it is a package for analyzing small molecules in R. It provides the scope for efficient prediction of physical and chemical properties of artificially synthesized molecules. It is capable of searching and classifying compounds and making similarity clusters (<https://www.bioconductor.org/packages/devel/bioc/vignettes/ChemmineR/inst/doc/ChemmineR.html>). The package comes with the ease of molecular visualization [42].
- "Cluster"—the "Cluster" package computes agglomerative hierarchical clustering of the dataset. Hierarchical agglomerative clustering is performed by merging smaller clusters on each step until one single large cluster exists. It determines agglomerative coefficient, which is a measure of the amount of clustering [43].
- "rgl"—"rgl" provides medium- to high-level function for 3D interactive graphics visualization. It also supports standard 2D image format like PNG and PGF [44].

```

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

> library(rodck)
Loading required package: rodcklibs
Loading required package: rJava
> library(chemometrics)
Loading required package: kpart
> library(ChemmineR)
> library(Cluster)
> library(rgl)
> library(vegan)
Loading required package: permute
Loading required package: lattice
This is vegan 2.5-6
> library(factoextra)
+
Loading required package: ggplot2
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBA
> library(fingerprint)
Attaching package: 'fingerprint'
The following object is masked from 'package:ChemmineR':
  fold

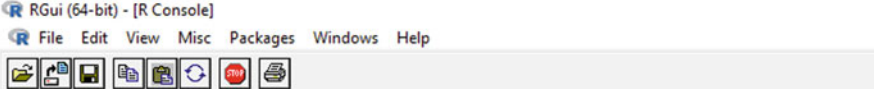
> library(fmcsR)
> library(factoextra)
> library(Nbclust)
Error in library(Nbclust) : there is no package called 'Nbclust'
> library(Nbclust)
> sdf.visualize(sdfset[1:4])
Error in sdf.visualize(sdfset[1:4]) : object 'sdfset' not found
> cdk.version()
[1] "2.3"
> data.frame()
data frame with 0 columns and 0 rows
> molecr<-load.molecules("substances.sdf")
> depictor <- get.depictor(style='ocb', abbr='reagents', width=300, height=300)

```

Fig. 21.8 R 4.0.2 showing response to loaded libraries. The program shows the version of the software being used

- “vegan”—this package provides a plethora of multivariate tools for diversity analysis, dissimilarity analysis, etc. Originally developed for analysis of descriptive community ecology, its tools can be used for many other types of data [45].
- “factoextra”—it can extract and visualize multivariate data analyzed, like principle component analysis (PCA), correspondence analysis (CA), and hierarchical multiple factor analysis (HMFA). “factoextra” R packages include “factomineR”, “ade4”, “stat”, “ca”, “mass”, and “exposition”. It facilitates clustering analysis and visualization. It also produces ggplot2 based on elegant data visualization [46].
- “fingerprint”—it contains functions to manipulate binary fingerprints of arbitrary length. Fingerprints can be converted to Euclidean vectors (i.e., points on the unit hypersphere) and can also be folded using OR. Arbitrary fingerprint formats can be handled via line handlers. Currently, handlers are provided for CDK, MOE, and BCI fingerprint data [39].
- “fmcsR”—it provides an R interface, with the time-consuming steps of FMCS algorithm implemented in C++. It includes utilities for pairwise comparisons, structure similarity searching, and clustering and visualization of maximum common substructures (MCSs). In comparison to an existing MCS tool, fmcsR shows better performance over a wide range of compound sizes [47] though its installation is not absolutely essential.

After all the packages have been included in the R library, “user-defined dataset” can be obtained using the “load.molecule” function (Fig. 21.8). There already exists an instance of dataset called the “bpdata” that can be used for practice. The bpdata contains 277 molecules which include SMILES and their



```

> bpdata

```

	SMILES	BP
bromo-trichloro-methane	C(Br)(Cl)(Cl)Cl	378.0
chloro-trifluoro-methane	ClC(F)(F)F	191.7
carbon tetrachloride	C(Cl)(Cl)(Cl)Cl	349.8
tetrafluoromethane	C(F)(F)(F)F	145.1
bromoform	BrC(Br)Br	422.3
chloro-difluoro-methane	C(Cl)(F)F	232.3
dichloro-fluoro-methane	C(Cl)(Cl)F	282.0
chloroform	C(Cl)(Cl)Cl	334.3
fluoroform	C(F)(F)F	191.0
dibromomethane	C(Br)Br	370.1
dichloromethane	C(Cl)Cl	312.9
difluoromethane	C(F)F	221.5
diiodomethane	C(I)I	455.2
formaldehyde	O=C	254.0
formic acid	C(=O)O	373.7
bromomethane	CBr	276.7
chloromethane	CCl	248.9
fluoromethane	CF	194.8
iodomethane	CI	315.6
methanol	CO	337.8
methanethiol	CS	279.1
1,1-dichloro-1,2,2-tetrafluoro-ethane	C(C(F)(F)F)(Cl)(Cl)F	276.2
1,1,1,2,2,2-hexafluoroethane	C(C(F)(F)F)(F)(F)F	194.9
2,2-dichloro-1,1,1-trifluoro-ethane	C(C(Cl)Cl)(F)(F)F	301.0
1,1,2-trichloroethylene	C(=C(Cl)Cl)Cl	360.1
2,2,2-trichloroacetaldehyde	C(C=O)(Cl)(Cl)Cl	370.8
1,1,1,2,2-pentachloroethane	C(C(Cl)(Cl)Cl)(Cl)Cl	433.0
1,1,1,2,2-pentafluoroethane	C(C(F)(F)F)(F)F	225.1
1,1,2,2-tetrabromoethane	C(C(Br)Br)(Br)Br	516.7
1,1,1,2-tetrachloroethane	C(CCl)(Cl)(Cl)Cl	403.7
1,1,2,2-tetrachloroethane	C(C(Cl)Cl)(Cl)Cl	418.3
1,1-difluoroethylene	C=C(F)F	187.5
1,1,2,2-tetrafluoroethane	C(C(F)F)(F)F	250.1
bromoethylene	C=CBr	288.9
acetyl chloride	CC(=O)Cl	323.9
1,1,1-trichloroethane	CC(Cl)(Cl)Cl	347.2

Fig. 21.9 Visualization of bpdata: List of molecules and their SMILES and boiling points in Kelvin

corresponding boiling points in Kelvin. By giving “bpdata” command on R console and pressing ENTER, one can look at the whole list (Fig. 21.9). These molecules can be viewed in 2D structure using the command:

```
>View.molecule.2d(name of the dataset[[entry no. of list]])
```

Choice of visualization can be obtained using the function “get.depictor”. The depictor function allows the choice of color of atom groups and depiction of functional groups:

```
>Depictor<-get.depictor(style='cob',abbr='reagents',width=300,
height=300)
```

```

> library(fmcsR)
> library(factoextra)
> library(Nbclust)
Error in library(Nbclust) : there is no package called 'Nbclust'
> library(NbClust)
> sdf.visualize(sdfset[1:4])
Error in sdf.visualize(sdfset[1:4]) : object 'sdfset' not found
> cdk.version()
[1] "2.3"
> data.frame()
data frame with 0 columns and 0 rows
> moleq<-load.molecules("substances.sdf")
> depictor <- get.depictor(style='cob', abbr='reagents', width=300, height=300)
>
> view.molecule.2d(moleq[[5]], depictor=depictor)
Error in .jnew("org/kuha/rcdk/view/ViewMolecule2D", molecule, as.integer(width), :
  java.lang.NoSuchMethodError: <init>
> data(bpdata)
> mols <- parse.smiles(bpdata[,1])
> fps <- lapply(mols, get.fingerprint, type='circular')

```

Fig. 21.10 Problem encountered during 2D visualization of molecules with RCDK package on Windows platform

```

A > data(bpdata)
> mols <- parse.smiles(bpdata[,1])
> fps <- lapply(mols, get.fingerprint, type='circular')
> fp.sim <- fingerprint::fp.sim.matrix(fps, method='tanimoto')
> fp.dist <- 1 - fp.sim
> cls <- hclust(as.dist(fp.dist))
> plot(cls, main='A Clustering of the BP dataset', labels=FALSE)
> data(moleq)

```

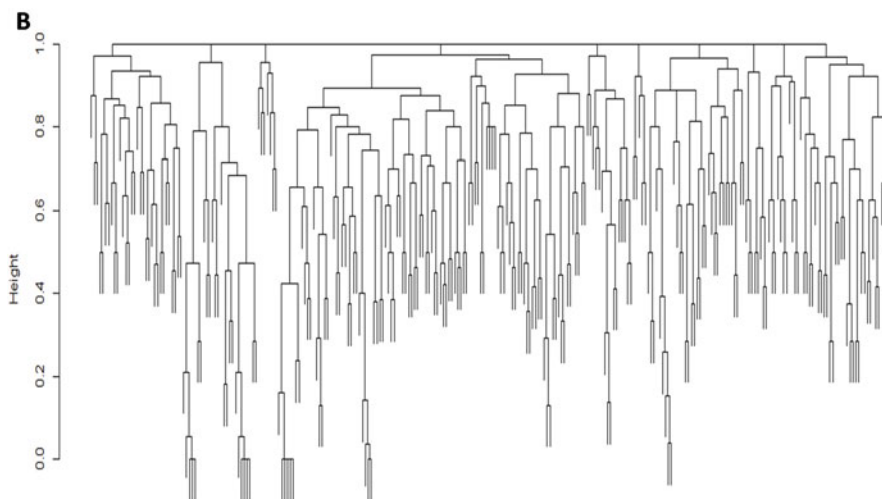


Fig. 21.11 (a) Generation of hierarchical clustering using Tanimoto matrix command. (b) Hierarchical clustering of molecules of BP dataset

```

A > query.mol <- parse.smiles('CC(=O)')[[1]]
> target.mols <- parse.smiles(bpdata[,1])
> query.fp <- get.fingerprint(query.mol, type='circular')
> target.fps <- lapply(target.mols, get.fingerprint, type='circular')
> sims <- data.frame(sim=do.call(rbind, lapply(target.fps,
+     fingerprint::distance,
+     fp2=query.fp, method='tanimoto')))
> subset(sims, sim >= 0.3)
      sim
C(=O)O  0.3333333
COC=O   0.3636364
CCC=O   0.3636364
CC(C)C=O 0.3636364
> |

B > query.mol <- parse.smiles('C(=O)O ')[[1]]
> target.mols <- parse.smiles(bpdata[,1])
> query.fp <- get.fingerprint(query.mol, type='circular')
> target.fps <- lapply(target.mols, get.fingerprint, type='circular')
> sims <- data.frame(sim=do.call(rbind, lapply(target.fps,
+     fingerprint::distance,
+     fp2=query.fp, method='tanimoto')))
> subset(sims, sim >= 0.3)
      sim
C(=O)O  1
> |

```

Fig. 21.12 (a) Similarity is found for depicted molecule. (b) No similarity is found

It is advisable to use `rcdk` on a Linux or macOS. Windows users often get an error message. This may be due to the fact that either `Rjava` is not installed or `R` cannot detect where `Rjava` is located. It is advisable that the user installed `Rjava` in the same directory in which `R` is located. Please refer to <https://github.com/CDK-R/cdkr/issues/61> for further discussion (Fig. 21.10).

The package can be utilized for determining Tanimoto similarity. Tanimoto coefficient is similar to that of Jaccard. The Tanimoto index is used to determine similarity between molecules. Without going deeper into the mathematics, the user can use this chemometrics tools under the `rcdk` package as long as the physical significance is clear. The program can generate hierarchal clustering using Tanimoto matrices. An example with the already available `bpdata` is shown in Fig. 21.11.

Similarity search can also be performed based on Tanimoto index with a user-defined cutoff. For example, here, two cases have been represented at which in the previous example molecules have been searched in the `bpdata` having Tanimoto similarity 0.3 or more for acetaldehyde. The result in output comes in the tabular form depicting the molecules and similarity index. When no similarity is obtained, the program generally returns the same molecule with Tanimoto index of 1. For example, formic acid similarity search does not provide any other molecular instances and hence returns formic acid itself (Fig. 21.12). Using molecular docking,

novel protein-binding molecules can be designed. Once the lead structure is obtained based on Tanimoto similarity, other analogues can be searched. Naturally available analogues can then be used to perform docking, and required derivatization/modification of the lead can be designed. Many other such functions and properties of molecules can be determined using the rcdk package. The users can use appropriate descriptor filter to perform their analysis of choice. Hence, as proficiency of the user increases in chemometrics, rcdk package can provide quality data by performing QSAR. Thus, it is advisable to explore this package in full form and take the opportunity of the diversity of R platform in the art of molecular analysis and drug design.

21.6 Conclusion

Investigating parasitic diseases like leishmaniasis through metabolomics has a huge scope and truly is a vast area to cover within the limited scope of this chapter. There are a lot of excellent literatures and reviews that are available to researchers and readers on metabolomics. However, the resources available are so vast that often, the problem is where to start with, especially for the beginners. The biologically significant outcomes of metabolomics data are often straightforward; however, the computational pipelines become major challenge even with good-quality data generated from well-designed experiments. Developing all skills of computational pipeline is a time-consuming process and a matter of intensive exploration and learning. With the advancement of technology as the cost of operating mass spectrometers becomes more and more affordable, a large volume of metabolomics data will be generated. Such data once freely available need mining; thus, computational basics are going to be a powerful skill that need to be developed and encouraged to obtain the maximum from already existing metabolomics data. We have tried to summarize some of these computational platforms that the readers can explore and try hands on without going into much details of programming. Thus, the focus should be more on understanding the algorithms rather than the programming language itself. We recommend R as the choice of learning statistical programming for metabolomics data analysis as it is user-friendly and backed by significant community support. Finally, it is always better to run different pipelines available for metabolomics on personal workstations as it provides the flexibility of data jiggling and understanding these platforms through data analysis and problem-solving.

Acknowledgment AV acknowledges the Department of Science and Technology, Government of India, for providing full financial support to conduct the project entitled "Identification of host-parasite interaction based on metabolomics & development of novel metabolic drug targets." AV also acknowledges full support of Dr. Sushil K. Jha, School of Life Sciences, Jawaharlal Nehru University, for conducting this project. This work was also supported to CKM by the Department of Science and Technology, India (grant no. VI-D&P/569/2016-17/TDT/C).

References

1. WHO Leishmaniasis (2020) World Health Organization Weekly Epidemiologic Record (WER) Global leishmaniasis surveillance, 2017–2018, and first report on 5 additional indicators. WHO 2020 WHO Fact Sheet 265–280
2. Steverding D (2017) The history of leishmaniasis. *Parasites Vectors* 10:82
3. Handman E, Bullen DVR (2002) Interaction of *Leishmania* with the host macrophage. *Trends Parasitol* 18:332–334
4. Olszewski KL, Morrisey JM, Wilinski D, Burns JM, Vaidya AB, Rabinowitz JD, Llinás M (2009) Host-parasite interactions revealed by *Plasmodium falciparum* metabolomics. *Cell Host Microbe* 5:191–199
5. Metallo CM, Vander Heiden MG (2013) Understanding metabolic regulation and its influence on cell physiology. *Mol Cell* 49:388–398
6. Chavali AK, Whittemore JD, Eddy JA, Williams KT, Papin JA (2008) Systems analysis of metabolism in the pathogenic trypanosomatid *Leishmania major*. *Mol Syst Biol* 4:177
7. Arjmand M, Madrakian A, Khalili G, Najafi Dastnaee A, Zamani Z, Akbari Z (2016) Metabolomics-based study of logarithmic and stationary phases of promastigotes in *Leishmania major* by 1H NMR spectroscopy. *Iran Biomed J* 20:77–83
8. Westrop GD, Williams RA, Wang L, Zhang T, Watson DG, Silva AM, Coombs GH (2015) Metabolomic analyses of *Leishmania* reveal multiple species differences and large differences in amino acid metabolism. *PLoS One* 10:e0136891
9. Vincent IM, Weidt S, Rivas L, Burgess K, Smith TK, Ouellette M (2014) Untargeted metabolomic analysis of miltefosine action in *Leishmania infantum* reveals changes to the internal lipid metabolism. *Int J Parasitol Drugs Drug Resist* 4:20–27
10. Pountain AW, Barrett MP (2019) Untargeted metabolomics to understand the basis of phenotypic differences in amphotericin B-resistant *Leishmania* parasites. *Wellcome Open Res* 4:176–176
11. Canuto GAB, Dörr F, Lago JHG, Tempone AG, Pinto E, Pimenta DC, Farah JPS, Alves MJM, Tavares MFM (2017) New insights into the mechanistic action of methyldehydrodieugenol B towards *Leishmania (L.) infantum* via a multiplatform based untargeted metabolomics approach. *Metabolomics* 13:56
12. Fernández-García M, Rojo D, Rey-Stolle F, García A, Barbas C (2018) Metabolomic-based methods in diagnosis and monitoring infection progression. In: Silvestre R, Torrado E (eds) *Metabolic interaction in infection*. Springer, Cham, pp 283–315
13. Verzani J (2011) Getting started with RStudio: “O’Reilly Media, Inc.” [Google Scholar]
14. Wickham H, Hester J, Chang W, Hester MJ (2020) Package ‘devtools’ [Google Scholar]
15. Ooms J (2014) The jsonlite package: a practical and consistent mapping between json data and r objects. arXiv preprint arXiv:14032805
16. Olivoto T, Lúcio ADC (2020) metan: an R package for multi-environment trial analysis. *Methods Ecol Evol* 11:783–789
17. Rinker T, Kurkiewicz D, Hughitt K, Wang A, Aden-Buie G, Burk L (2019) pacman: package management tool. R package version 05, 1
18. Pollard KS, Gilbert HN, Ge Y, Taylor S, Dudoit S (2020) Package ‘multtest’. *Bioconductor* <https://www.bioconductor.org/packages/release/bioc/manuals/multtest/man/multtest.pdf>
19. Gentleman R, Carey V, Huber W, Hahne F (2020) Package ‘genefilter’. *Bioconductor* <http://bioconductor.org/packages/release/bioc/manuals/genefilter/man/genefilter.pdf>
20. Goeman J, Oosting J, Finos L, Solari A, Edelmenn D (2020) The Global Test and the globaltest R package. *Bioconductor* <http://www.bioconductor.org/packages/release/bioc/vignettes/globaltest/inst/doc/GlobalTest.pdf>
21. Sharma G, Martin J (2009) MATLAB®: a language for parallel computing. *Int J Parallel Prog* 37:3–36

22. Libiseller G, Dvorzak M, Kleb U, Gander E, Eisenberg T, Madeo F, Neumann S, Trausinger G, Sinner F, Pieber T (2015) IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinform* 16:118
23. Xia J, Wishart DS (2016) Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. *Curr Protoc Bioinf* 55:14.10.11–14.10.91
24. Karger DR, Quan D (2004) Haystack: a user interface for creating, browsing, and organizing arbitrary semistructured information. In *CHI'04 extended abstracts on Human factors in computing systems*, pp. 777–778
25. Olivon F, Grelier G, Roussi F, Litaudon M, Touboul D (2017) MZmine 2 data-preprocessing to enhance molecular networking reliability. *Anal Chem* 89:7836–7840
26. Lei Z, Li H, Chang J, Zhao PX, Sumner LW (2012) MET-IDEA version 2.06; improved efficiency and additional functions for mass spectrometry-based metabolomics data processing. *Metabolomics* 8:105–110
27. Gatto L, Gibb S, Rainer J (2021) MSnbase, efficient and elegant R-based processing and visualization of raw mass spectrometry data. *J Proteome Res* 20(1):1063–1069
28. Gatto L, Gibb S (2016) MSnbase: labelled and label-free MS2 data pre-processing [Google Scholar]
29. Rainer J, Witting M, Neumann S (2020) LC-MS/MS data analysis with xcms [Google Scholar]
30. Gibb S (2019) MALDIquantForeign: Import/Export routines for MALDIquant. <https://CRAN.R-project.org/package=MALDIquantForeign>
31. Dieterle F, Ross A, Schlotterbeck G, Senn H (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem* 78:4281–4290
32. Gibb S, Strimmer K (2012) MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics* 28:2270–2271
33. Ryan PB, Soczek ML, Treitman RD, Spengler JD, Billick IH (1988) The Boston residential NO₂ characterization study—II. Survey methodology and population concentration estimates. *Atmos Environ* (1967) 22:2115–2125
34. King RL, Ruffin C, LaMastus F, Shaw D (1999) The analysis of hyperspectral data using Savitzky-Golay filtering—practical issues. 2. In *IEEE 1999 international geoscience and remote sensing symposium. IGARSS'99* (Cat. No. 99CH36293), vol 1, pp 398–400
35. Gibb S, Strimmer K (2011) Analysis of proteomics data using MALDIquant. In *Proceedings of the 8th international workshop on computational systems biology, WCSB*, pp. 49–52
36. Aversch NJ, Krömer JO (2018) Metabolic engineering of the shikimate pathway for production of aromatics and derived compounds—present and future strain construction strategies. *Front Bioeng Biotechnol* 6:32
37. Ji R-R, Ott K-H, Yordanova R, Bruccoleri RE (2011) FDR-FET: an optimizing gene set enrichment analysis method. *Adv Appl Bioinf Chem* 4:37
38. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, DuGar B, Feldstein AE, Britt EB, Fu X, Chung Y-M (2011) Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* 472:57–63
39. Guha R (2007) Chemical informatics functionality in R. *J Stat Softw* 18:1–16
40. Wehrens R (2011) *Chemometrics with R: multivariate data analysis in the natural sciences and life sciences*. Springer Science & Business Media
41. Buttrey SE, Karo C (2002) Using k-nearest-neighbor classification in the leaves of a tree. *Comput Stat Data Anal* 40:27–37
42. Backman TW, Cao Y, Girke T (2011) ChemMine tools: an online service for analyzing and clustering small molecules. *Nucleic Acids Res* 39:W486–W491
43. Kassambara A (2017) *Practical guide to cluster analysis in R: unsupervised machine learning*, vol 1: Sthda

44. Adler D, Nenadic O, Zucchini W (2003) Rgl: a r-library for 3d visualization with opengl. In Proceedings of the 35th symposium of the interface: computing science and statistics, Salt Lake City, vol 35, pp 1–11
45. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'hara R, Simpson GL, Solymos P (2016) vegan: community ecology package. R package version 2.4-3. Vienna: R foundation for statistical computing [Google Scholar]
46. Kassambara A, Mundt F (2017) Package 'factoextra'. Extract and visualize the results of multivariate data analyses, 76 [Google Scholar]
47. Wang Y, Backman TW, Horan K, Girke T (2013) fmcsR: mismatch tolerant maximum common substructure searching in R. *Bioinformatics* 29:2792–2794



Metabolomics: A Promising Tool to Study Disease Biomarkers and Host-Pathogen Interactions

22

Megha, Preeti, and Tulika Prasad

Abstract

Metabolomics is a comprehensive analysis of small-molecule metabolite profiles of cellular processes using mass spectrometry combined with advanced techniques of gas/liquid chromatography and nuclear magnetic resonance spectrometry. Whenever microbial pathogens invade the host cellular system, the host immune system responds via several innate and adaptive mechanisms to eliminate the pathogens. As a result, both host and microbial pathogens elicit adaptive responses, leading to altered metabolic pathways such as glycolysis, fatty acid, amino acid biosynthesis, and thereby cellular metabolites. Finally, either host cells or pathogens survive. These unique metabolites are analyzed by metabolomics and used as biomarkers for disease and pathogen detection, differentiating between microbial pathogens, prediction of altered metabolic pathways during infection, and development of drug resistance. Metabolomics completes the loop of central dogma beyond the proteins and provides the information on molecular phenotyping. Metabolomics has emerged as a promising tool to aid better understanding of metabolic pathways and new drug targets, rapid disease diagnosis, and development of effective, shorter, personalized therapy. Herein, the application of metabolomics in clinical setting for the study of infectious diseases, viz., pathogen and disease detection, differentiating between microbial pathogens, alterations in associated biochemical pathways during infection, and drug resistance, is discussed. This chapter highlights the analytical techniques used for metabolite profiling.

Megha · Preeti · T. Prasad (✉)

Special Centre for Nano Sciences and AIRF, Jawaharlal Nehru University, New Delhi, India

e-mail: prasadtulika@mail.jnu.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_22

403

Keywords

Metabolomics · Microbial infections · Metabolites · Mass spectrometry · NMR · Microbial adaptation · Diagnosis

22.1 Introduction

Living cell is itself a perfect machine, where coded information on DNA is transcribed into mRNA and then translated to proteins, which further initiate controlled and functional metabolic processes of the cell. Omics is the systematic study of biological interaction in cells such as genes, transcripts, protein, and metabolites. Genomics, transcriptomics, and proteomics mostly provide the information on genome and its product but convey very limited information about phenotype. Interestingly, compounds with low molecular weight (less than 1 KDa) or metabolites of cellular processes are closely linked to molecular phenotype, where metabolites are the functional representation of cellular state and their abundance is often related to homeostasis dysregulation or disease occurrence [1]. The entire set of metabolites within biofluids, cells, tissues, and organisms are termed as metabolome, and their identification and quantification is called metabolomics. Metabolomics is a powerful tool for clinical diagnosis of diseases, to measure chemical phenotypes, and to provide highly integrated profiles of biological status of cells [1]. Mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy are highly reproducible, quantitative techniques which are routinely used for the study of metabolites. However, other techniques such as capillary electrophoresis, infrared spectroscopy, and Raman spectroscopy are also used to study several metabolic disorders [2].

A recent metabolomics study reported the differential abundance of 373 and 204 metabolites within COVID-19 patients, indicating the disease severity of the COVID-19 pandemic [3]. The use of metabolomics in screening of infectious diseases has shown accurate and noninvasive diagnosis of metabolic biomarkers associated with microbial infections and diseases [4]. This technique has been widely used for the extensive study of several common diseases like *Mycobacterium tuberculosis* (*M. tuberculosis*), human immunodeficiency virus infection (HIV), hepatitis C, etc. [4]. Metabolomics can be used for detection/identification of disease and pathogen, to differentiate between microbial strains, to provide the information on virulence factors, for adaptation features of microbes in host environment, and to identify the novel compounds produced by microbes. The comprehensive characterization of metabolic phenotypes can progress precision medicine to various levels, including the characterization of metabolic derangements underlying the disease, discovery of new therapeutic targets, and discovery of biomarkers that may be used for either disease diagnosis or monitoring activity of therapeutics. Herein this

chapter, we have summarized the metabolomics techniques and their applications to study disease biomarkers and host-microbial pathogen interactions.

22.2 Analytical Techniques for Profiling of Metabolites

MS and NMR spectroscopy are complementary analytical techniques used for qualitative and quantitative analysis of metabolites extracted from various biological samples or biofluids (plasma, urine, blood, breath) of healthy and diseased individuals. These are highly sensitive, powerful, high-throughput techniques, capable of identifying several hundred metabolites in a single measurement and helping to understand the differences in the biological pathways of healthy and diseased individuals. Figure 22.1 depicts the general workflow for metabolomics.

22.2.1 Mass Spectrometry (MS)

MS measures the mass-to-charge (m/z) ratio of different molecules within a sample and calculates the exact molecular weight of components present in the sample. MS quantifies both known and unknown compounds within a sample via molecular weight determination and helps in elucidation of the chemical structure and, thus, chemical properties of different molecules. It is an ultrasensitive technique, which can detect molecules with high resolution ($\sim 10^3$ – 10^4), at even very low concentrations of femtomolar (10^{-15} MolesL $^{-1}$) to attomolar (10^{-18} MolesL $^{-1}$), and routinely used to analyze hundreds of compounds in a single sample [5]. MS is successfully used in metabolomics and provides new perspective to understand the

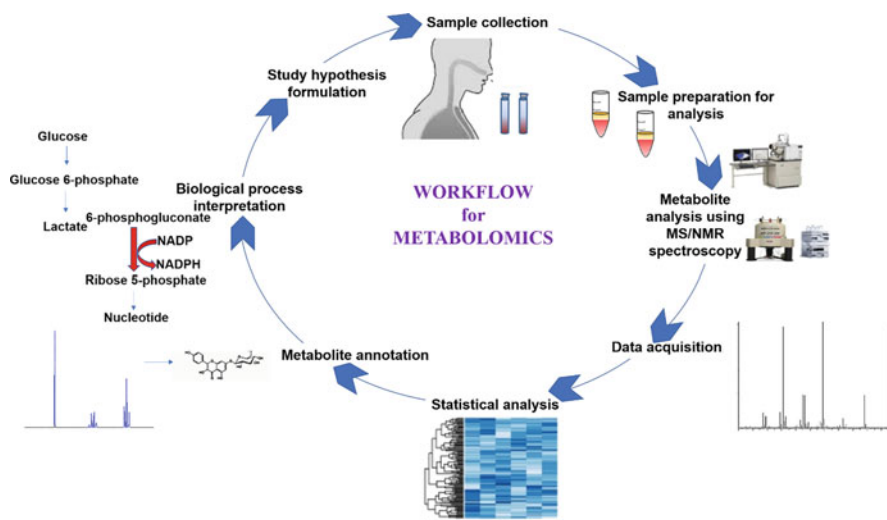


Fig. 22.1 General workflow for metabolomics

cellular processes, alterations in biochemical pathways due to host-pathogen interactions, and identification of disease biomarkers [6, 7]. MS typically consists of three functional units, namely, ionization source, mass analyzer, and ion detection system. The ionization source is used to ionize the analyte to gas phase ions, while the mass analyzer further sorts and separates the ions based on mass-to-charge ratio. After that, the separated ions are measured by the detector, and mass spectrum is generated based on m/z ratio of ions in sample and their intensities. Electrospray ionization (ESI) and matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) are two ionization techniques, majorly used in metabolomics because of their soft nature of ionization, which prevents the fragmentation of molecular ions [5]. MALDI-TOF is routinely used for the identification of microbial strains. In MALDI-TOF, the cultured microbial colony from the collected sample is first transferred onto a MALDI target spot and then embedded in a matrix, and subsequently a laser is aimed at the target spot [8]. The spectrum generated requires further data processing before comparing with the reference database of known microorganisms. The National Institute of Standards and Technology (NIST) MS Data Center develops and makes available the database of reference GC/LC-MS libraries [9] and related software tools for identification of compounds by comparison with reference mass spectra. This makes it convenient for detection of microbial strains and also helps in elucidation of mechanisms involved in microbial pathogenesis [10, 11]. The Bruker Biotyper and Vitek MS bioMérieux are FDA-approved platforms for rapid, cost-effective, and accurate identification of microbes [12].

However, MS most often requires pure sample for identification; therefore combining MS with chromatographic separation techniques maximize the ability of MS for accurate identification of compounds. Gas chromatography (GC) and liquid chromatography (LC) are pre-MS separation methods, wherein GC is used for the separation of volatile compounds, while thermally unstable and nonvolatile molecules are separated by LC. More polar biomolecules such as organic acids, nucleosides, nucleotides, ionic species, polyamines, organic amines, etc. are analyzed by LC-MS, where mobile phase is liquid and no derivatization process is required [10, 11]. GC-MS is widely used for analysis of alcohols, esters, alkaloids, amino acids, sugars, drugs, toxins, fatty acids, etc. [10, 11]. For GC, prior to analysis, sample is chemically derivatized to convert the nonvolatile compounds to volatile form. Derivatization improves sample volatility, selectivity, detectability, separation, and thermal stability in chromatographic applications, increases retention time, enlarges substrate spectrum, and removes tailing. GC-MS is a robust analytical technique with high sensitivity, selectivity, and reproducibility. GC-MS overcomes the limitations of matrix effect and ion suppression associated with LC-MS by co-eluting the compounds, therefore resulting in greater chromatographic resolution [10, 11]. Interestingly, studies have demonstrated the successful use of GC-MS-based metabolomics for the identification of microbial strains and biomarkers associated with pathogenicity and infectious diseases [13]. In recent years, mass spectrometry imaging (MSI) has emerged as a promising technique which enables the in situ detection of several compounds, ranging from metabolites to proteins, and can also simultaneously provide spatiotemporal distribution of molecular species in

a variety of samples [10, 11, 14]. It has the capability to image thousands of molecules, such as metabolites, lipids, peptides, proteins, and glycans, in a single experiment without labeling. The combination of information gained from mass spectrometry (MS) and visualization of spatial distributions in thin sample sections makes this a valuable chemical analysis tool for biological specimen characterization. MSI is increasingly used as an analytical technique both as complement and replacement to other imaging methods. By improving sample preparation protocols, instrument throughput, resolution capabilities, streamlined data analysis, and quantitation, it is anticipated that MSI will be routinely utilized in clinical settings. Advances in MS has certainly improved the selectivity, sensitivity for detection at lower concentration, compatibility with separation techniques, and ability for qualitative and quantitative analysis, thus making MS an ideal analytical technique for profiling of metabolites in clinical settings.

22.2.2 Nuclear Magnetic Resonance (NMR)-Based Spectroscopy

NMR is one of the principal analytical techniques used in metabolomics, preferably for long-term or large-scale metabolite profiling, when it is not restricted to analysis of only biofluid or tissue extract. Being a nondestructive analytical technique, NMR is more advantageous to use for metabolomics. Moreover, NMR spectroscopy has relatively high reproducibility and high throughput and requires almost no separation techniques. The principle of NMR is based on the magnetic properties, referred to as “spin,” of certain atomic nuclei to provide information about their immediate environment. Each metabolite is made up of atom consisting of nuclei, and when magnetic field is applied, specific NMR signal is obtained by the radio-frequency pulses that interact with the nuclei of an atom and characterize the resonate frequency of that nuclei based on the environment and chemical surroundings [15]. Thus, the environment of that nucleus, the presence of electrons and protons on neighboring atoms, and their interactions affect the magnetic field, and thus, the energy required to flip the nucleus. Metabolites such as alcohols, sugars, highly polar compounds, organic acids, etc. are detected and characterized by NMR spectroscopy [15]. Moreover, different classes of metabolites such as nitrogen-containing, phosphorous-containing, and protein-bound metabolites (lipoprotein particles) and certain inorganic metabolites (H^+ ions and metal ions) can be identified either separately or simultaneously [15, 16]. MS-based analytical methods being destructive, both LC-MS and GC-MS are unsuitable for analyzing living samples, but with NMR spectroscopy, the metabolite profiling of living cell in real time is possible [17, 18]. Milner et al. used 1H NMR for the analysis of metabolites in urine and feces samples of healthy and influenza virus-infected mice. Significantly elevated levels of certain metabolites such as acetylcarnitine, ascorbate, glucose, and 3-hydroxybutyrate were found in urine sample of obese mice indicating alterations in metabolic pathways associated with kidney [19]. Similarly, discriminant metabolites such as acetate and trimethylamine were rapidly identified by 1H NMR spectroscopy during urinary tract infection caused by *Escherichia coli*

Table 22.1 Comparison between MS and NMR

Mass spectrometry	NMR spectroscopy
Needs derivatized pre-sample preparation for GC-MS analysis	Easy or minimal sample preparation with no need of derivatization step and has high reproducibility
MS is not suitable for profiling of living cell metabolites	Suitable for profiling of living cell metabolites [22]
For maximization of metabolites detection, MS requires different ionization methods	Rapid analysis, all metabolites can be observed in a single measurement
Highly sensitive and selective method. It is suitable for targeted analysis, can operate for sample even at concentration of 10^{-6} M, and can be efficient up to 10^{-18} M [23]	Almost 10–100 times less sensitive than GC-MS/LC-MS but suitable for both targeted and untargeted analysis and able to measure sample between concentration 10^{-4} to 10^{-5} M [23]
Ionization is an important factor for resulting data. MS line intensity is not directly correlated to metabolite concentration [15]	Easy quantitative method, signal intensity is directly proportional to number of nuclei in the molecule and to metabolite concentration [15]
MS could be used for metabolic flux analysis but destructive nature of MS-based methods resulted in more limitations than NMR-based fluxomics	Suitable for in vivo and in vitro metabolic flux analysis [24, 25]

(*E. coli*) [20]. With simultaneous use of magnetic resonance spectroscopy (MRS) and magnetic resonance imaging (MRI), NMR can be a versatile tool in biomedical field for metabolite imaging and analysis of living samples [21]. Table 22.1 gives a comparison between MS and NMR.

22.3 Metabolomics for Infectious Diseases

Human beings are under continuous exposure to various pathogenic and nonpathogenic microorganisms such as viruses, bacteria, parasites, and other extrinsic factors. The human gut itself harbors a complex and dynamic population of microorganisms called the gut microbiota, comprising beneficial, commensal, and opportunistic microbes, which exert a remarkable influence on the host during homeostasis and disease. The gut microflora plays a significant role in host nutrient metabolism, xenobiotic and drug metabolism, maintenance of structural integrity of the gut mucosal barrier, immunomodulation, and protection against pathogens. Although the microorganisms which are part of gut microflora are well adapted to the human body, under certain unfavorable and immunocompromised conditions, few commensal microorganisms often turn pathogenic to result in disease such as candidiasis caused by the opportunistic fungus, *Candida albicans*. On the other hand, there are many other microbes which are not part of the human microbiota and can cause several diseases. Some common examples of infectious diseases are urinary tract infections, common cold, tuberculosis, HIV/AIDS, influenza, viral hepatitis,

measles, typhoid, chicken pox, diphtheria, dengue, chikungunya, etc.; however, the list remains exhaustive.

Infectious diseases are major causes of mortality and morbidity worldwide. They are transmissible and caused by various pathogens, including bacteria, fungi, viruses, and parasites. Infections by these pathogens result in illness and disease. Signs and symptoms and treatment of infectious diseases depend on the host and the pathogen. The invasion of host by the microbial pathogens and pathogenicity have been widely studied since decades, and several vaccines or therapeutics have been developed to limit the microbial infections. While vaccines are available as protective measures for many diseases but for others, vaccines are still not available. Treatments for diseases are dependent on therapeutic targets that are already available or still not known yet. Even in this era of ongoing COVID-19 pandemic which has severely impacted several millions of the world population and resulted in millions of deaths in 2020 (<https://www.worldometers.info/coronavirus/>), there is still no vaccine or medicine currently available for COVID-19. However, many of the symptoms can be treated, and getting early care from a healthcare provider can make the disease less dangerous. Undoubtedly, COVID-19 has led to a drastic change in the social structure and lifestyle of humans globally. Therefore, in the light of such background, it is essential to identify disease at early stage and also to find therapeutic targets to treat infectious diseases. The design and development of vaccines and therapeutic drugs depend on many factors including the host immune system and type of the pathogen.

As a part of adaptation to the surrounding environment and protection from microbial pathogens, the immune system plays a vital role in providing protection to the human host. Human immune system consists of innate and adaptive immune systems. Soon after infecting a human host, the microbial pathogen tries to hijack the host cellular system for its survival. Existing literature supports the fact that the host metabolic pathways get perturbed during infection, and in order to protect against infection, integrins, cytokines, and tissue-specific immune cells such as monocytes, macrophages, B cells, and T cells are activated to destroy the microbial pathogen. Therefore, it is essential to identify the metabolites, and the holistic approach of systems biology used for metabolite profiling is called “metabolomics.” Systems biology integrates omics technologies such as genomics, transcriptomics, proteomics, and metabolomics. Of these, metabolomics is relatively a novel approach and offers a robust platform for researchers to understand the influence of different factors in the metabolic pathway of an infected cell. From a clinical standpoint, metabolomics can detect the pattern of metabolites that are associated with particular disease or pathogen, may detect an unknown metabolite indicating new pathophysiological conditions, and also trace the effectiveness or toxicity of a drug in disease. This would certainly aid in early detection of pathogen and disease, facilitate timely disease prevention and treatment, identify new drug targets, and improve therapeutic interventions.

Identification of microbial species is a crucial bottleneck for clinical diagnosis of infectious diseases. Quick and reliable identification is a key factor to provide suitable antimicrobial therapies and avoid the development of multiple-drug

resistance. The conventional methods for identification of microbial pathogens are very tedious and time-consuming and require laboratory skills and proper clinical setup. This delays the identification of pathogen and early disease diagnosis and thus, impedes timely treatment of disease. However, application of metabolomics for clinical disease diagnosis and identification of microbial pathogens certainly opens new prospects, and chemical analysis of microbial metabolites can facilitate rapid detection of the pathogens and help differentiate between them.

22.3.1 Identifying and Differentiating Between Microbial Pathogens

Volatile organic compounds (VOCs) are low molecular weight and carbon-based organic molecules with vapor pressure of ≥ 0.01 kPa at 20 °C and are naturally volatile in ambient temperature [26]. In 1964, coliform bacteria were detected via profiling of VOCs which lead to identification of indole, acetoin, pyruvate, and 2,3-butanediol in culture media [27]. VOCs are produced by resident microbes and various other microbes as primary (e.g., ethanol, acetone, acetic acid) or secondary metabolites (signaling molecules) (Table 22.2) [26]. These VOCs can diffuse into breath, urine, feces, and sweat; therefore, analysis of VOCs involves an inexpensive, noninvasive method for collection of samples. Profiling of VOCs can help fingerprint metabolites produced by the infecting pathogen or reflect pathogen-induced host responses or a combination of both (Fig. 22.2). Testing for volatile biomarkers in clinical samples offers an option for developing rapid and potentially inexpensive disease screening tools. The testing of volatiles can be performed frequently in follow-up studies, which may indicate disease progression and be helpful in monitoring therapeutic intervention. The direct link of diffused VOCs to microbial infection in host was well demonstrated by specific VOCs released by the resident pathogen during host-pathogen interactions [34], e.g., indole metabolite found in *E. coli* infections [35], 2-aminoacetophenone compound reported in infection by the respiratory tract pathogen, *Pseudomonas aeruginosa* (*P. aeruginosa*) [36], and high level of *p*-menth-1-en-8-ol metabolite detected in fecal samples of cholera patients infected with *Vibrio cholerae* (*V. cholerae*) (Table 22.2) [32]. VOCs collected from exhaled breath require noninvasive, less strenuous sampling and are widely diagnosed by GC-MS and NMR spectroscopy. However, ion-molecule reaction-mass spectrometry (IMR-MS), field asymmetric ion mobility spectrometry (FAIMS), and selected ion flow tube-mass spectrometry (SIFT-MS) are also used for profiling VOCs [37]. The profiling of VOCs showed 68% sensitivity and 100% specificity for the detection of *P. aeruginosa* in sputum and cough swab samples of cystic fibrosis patients [38], by detecting volatile hydrogen cyanide (HCN) using SIFT-MS [38, 39]. HCN may be a specific indicator of *P. aeruginosa* infection in vivo and offers promise as a biomarker for noninvasive detection of *P. aeruginosa* infection by breath analysis. In another study, it was possible to discriminate between healthy controls and cystic fibrosis patients with or without *P. aeruginosa* colonization based on detection of C5–C16 hydrocarbons and *N*-methyl-2-methylpropylamine, using gas chromatography-time of flight-MS [40]. Active

Table 22.2 Specific VOCs produced by various microbial pathogens

VOCs	Pathogen	Method	Source	Disease	Reference
Naphthalene, 1-methyl-cyclohexane, hexyl-heptane, 2,2,4,6,6-pentamethyl-benzene, 1,3,5-trimethyl-1-hexene, 4-methyl	<i>M. tuberculosis</i>	GC-MS	Breath	Active TB	[28]
HCN	<i>P. aeruginosa</i>	SIFT-MS	Breath	Cystic fibrosis	[29]
2-Aminoacetophenone		GC-MS			[30]
2,3-Butanediol, [R-(R*,R*)]-, hexadecane; and undecane, 3,8-dimethyl metabolites, <i>N,N</i> -dimethylacetamide; phosphonic acid, (<i>p</i> -hydroxyphenyl); 3,5-decadien-7-yne, 6- <i>t</i> -butyl-2,2,9,9-tetramethyl; 1,6-dioxacyclododecane-7,1,2-dione; caprolactam; 5,7-octadien-2-one, 3-acetyl; nonanal; and 5-hepten-2-one, 6-methyl-	<i>Helicobacter pylori</i>	GC-MS	Breath	Gastric cancer	[31]
Dimethyl disulfide	<i>Vibrio cholerae</i>	GC-MS	Feces	Cholera	[32]
<i>p</i> -Menth-1-en-8-ol					[32]
3-Methyl-2-butanone and styrene	<i>C. albicans</i>	GC-MS	Breath	Oral candidiasis	[7]
Combination of <i>p</i> -xylene, 2-octanone, 2-heptanone, <i>n</i> -butyl acetate	<i>C. krusei</i>	GC-MS			
1-Hexanol	<i>C. tropicalis</i>	GC-MS			
α -Trans-bergamotene, β -trans-bergamotene, a β -vaitrenene-like sesquiterpene, or trans-geranylacetone	<i>A. fumigatus</i>	GC-MS	Breath	Invasive aspergillosis	[33]

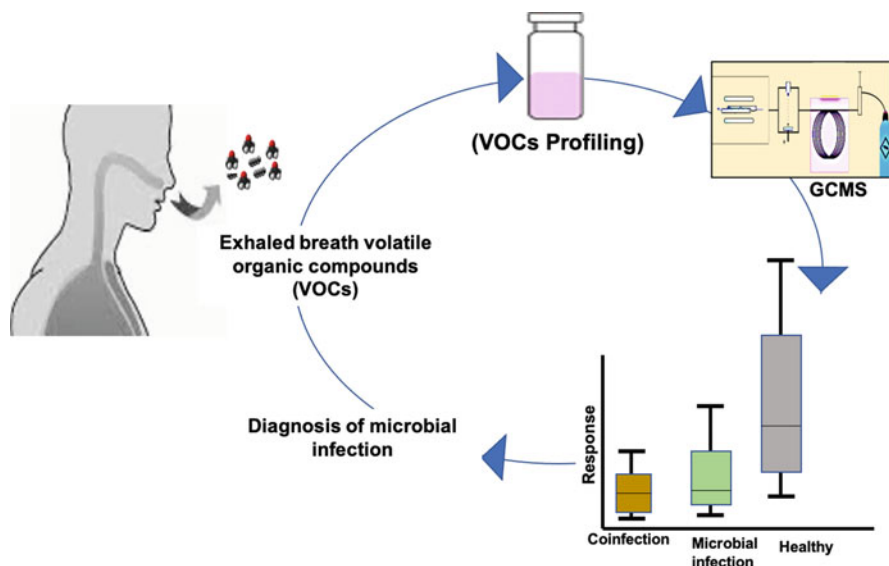


Fig. 22.2 Illustration for profiling of VOCs for identification of microbial pathogens and disease diagnosis

pulmonary tuberculosis (TB) was distinguished from nonactive TB by breath VOC patterns wherein 1,3,5-trimethylbenzene was identified in active pulmonary TB and 1,2,3,4-tetramethylbenzene in the nonactive stage [41]. Profiling VOCs in urine samples of TB patients and healthy controls using GC-MS coupled to a headspace sampler revealed a panel of five selected biomarkers, namely, alpha-xylene, isopropyl acetate, 3-pentanol, dimethylstyrene, and cymol, which enabled discrimination between TB-infected and healthy individuals with an accuracy of 98.8% (area under the curve [AUC] of 0.988) [41]. VOCs which were principally derivatives of naphthalene, benzene, and alkanes and metabolic products of *M. tuberculosis* were identified in breath samples of TB patients as biomarkers of pulmonary TB [28]. *Helicobacter pylori* (*H. pylori*) is the main etiological factor of gastritis and associated with duodenal ulcer and gastric cancer. Its diagnosis involves invasive method of endoscopy with gastric biopsies. Specific VOC profiles were detected for *H. pylori* by noninvasive expired air method, which is a low-cost method with good patient compliance and can be used for gastric cancer diagnosis [31]. Fungal pathogens were also identified based on signature volatiles such as 3-methyl-2-butanone and styrene and were detected as characteristic VOCs of *C. albicans* [7]. However, Koo et al. clearly distinguished invasive aspergillosis from pneumonia with 94% sensitivity and 93% specificity, wherein α -trans-bergamotene, β -trans-bergamotene, a β -vatiirenene-like sesquiterpene, or trans-geranylacetone VOCs (secondary metabolites) were identified in invasive aspergillosis patients (Table 22.2) [33]. Traxler et al. reported high concentrations of acetaldehyde, propanal, and n-propyl acetate in the breath of pigs which indicated coinfection by both bacteria and influenza A virus and an interaction between the pathogens (Table 22.2)

[42]. Electronic nose sensor (gas sensor array) is a rapid, accurate sensing technique, popular for microbial screening. Fend et al. used electronic nose sensor to detect *M. tuberculosis* in both culture and spiked sputum samples and achieved the detection limit of 1×10^4 mycobacteria ml^{-1} with specificity of 91% and sensitivity of 89% (Table 22.2) [43]. The ability of an electronic nose (e-nose) to detect *M. tuberculosis* in clinical specimens opens the way to developing this method as a rapid, automated system for early diagnosis of respiratory infections. Although e-nose sensors are much simpler and cost-effective, metabolomics techniques are able to provide details of pathophysiology of patient. However, VOCs collected from exhaled breath could be influenced by environment or activity before or during sampling; therefore, lack of consistency of proposed breath biomarkers is a certain limitation associated with profiling of VOCs [44]. Nevertheless, profiling VOCs still is a promising, novel, noninvasive, pathogen-specific approach for precise identification and monitoring of disease and discrimination of diseased person from healthy.

22.3.2 Fingerprinting Metabolic Differences in Biochemical Pathways of Host and Pathogen Associated with Host-Pathogen Interactions, Disease, Pathogenicity, and Drug Resistance

Fingerprinting metabolic differences between differentially regulated biochemical pathways of host and pathogen can facilitate the discovery of potential biomarkers associated with human dysbiosis or microbial diseases and provide insights into the host metabolites, microbial metabolites, their potential functions, impact on host-pathogen interactions, disease, pathogenesis, and drug resistance (Fig. 22.3).

Upon invasion of the host cells by the microbial pathogens (bacteria, virus, fungi, and protozoa), the pathogen-associated molecular patterns (PAMPs) present on microbes are recognized in host immune cells by pathogen recognition receptors (PRR) such as retinoic acid-inducible gene I-like receptors, C-type lectin receptors, and nucleotide-binding oligomerization domain-like receptors. Subsequently, host-defense mechanism is activated against the pathogen resulting in the production of antiviral agents, proinflammatory cytokines/chemokines and antibodies [45]. The host-pathogen interactions result in altered metabolic products (glucose, amino acid, lipids, and nucleotides) and synthesis of metabolites or several virulence factors (e.g., adhesins, modulins, toxins, etc.) in the microbial pathogens, which allow them to successfully replicate in host cell by evading host immune system [46]. Targeted metabolomics of the stationary phase growth arrested epimastigotes, and exponentially growing *Trypanosoma cruzi* parasites, known to cause Chagas disease, revealed the adaptive metabolic changes that epimastigotes undergo before they get into the metacyclic trypomastigote stage [47]. This finely tuned adaptive metabolic mechanism enables switch from highly reduced, energy-rich metabolites such as glucose to oxidized energy-poorer nutrients such as amino acids, found abundantly in the stationary phase of *T. cruzi* epimastigote. This metabolic plasticity might be crucial for the survival of the parasites under different environmental conditions. Knowledge of metabolic capabilities during their life cycle can reveal

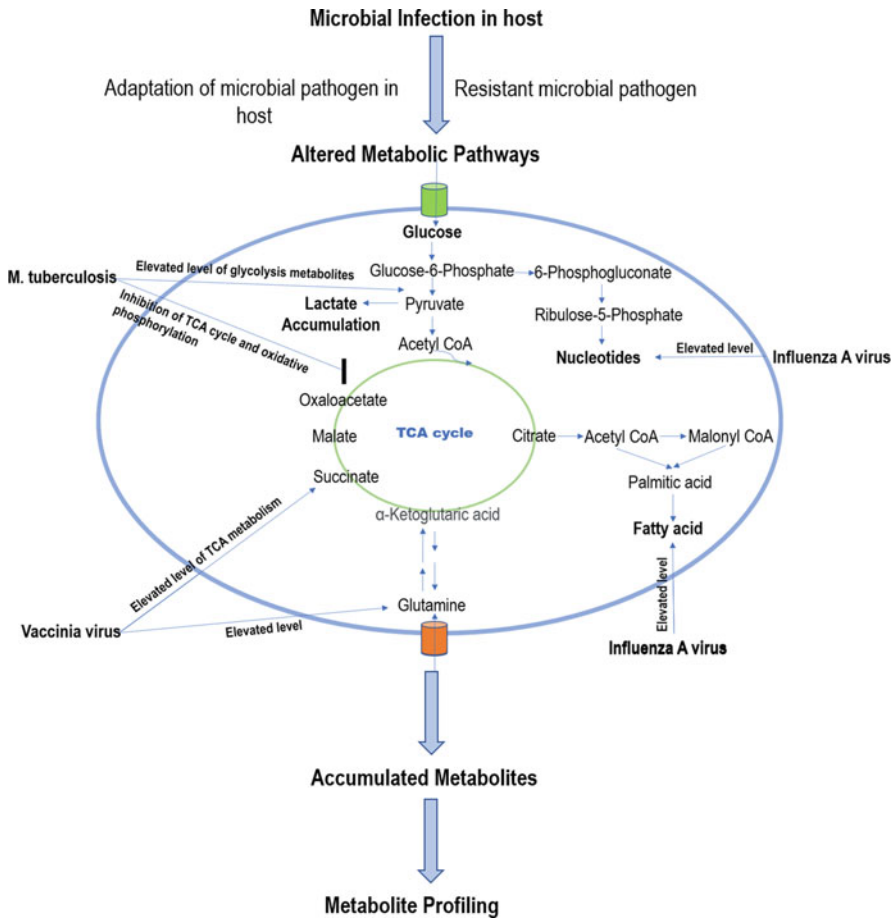


Fig. 22.3 Schematic representation for alterations in metabolic pathways, intermediates, and metabolites due to host-pathogen interactions

metabolic checkpoints as novel targets for designing therapeutic interventions and disease control. Triatomine is a vector for *T. cruzi* and during the replication and passage of *T. cruzi* through its intestinal tube, the vector immune system produces significant amount of oxidants [6]. It is suggested that triatomine recognition of PAMPs trigger innate and humoral immunity and cellular protection.

The Warburg-like metabolism (increased uptake of glucose and increased production of lactate in presence of oxygen) observed during infection by *Mycobacterium tuberculosis* is also an example of switching of bioenergetic metabolic pathways (switch from oxidative phosphorylation to glycolysis) [48]. Metabolic changes revealed that during infection, pyruvate is converted to lactate instead of entering into tricarboxylic acid (TCA) cycle and results in accumulation of intermediates of glycolysis and lactate, which are further used as an additional energy source for growth of bacteria (Fig. 22.3). Remarkably, the ability to use

glucose-6-phosphate and/or amino acids instead of glucose favors the activation of certain virulence factors or genes or metabolites, which are essential for the cytosolic lifestyle of the pathogen and help in the adaptation of bacteria to harsh environments such as the urinary tract, the blood stream, and the meninges [49]. There exists links between metabolism and expression of those genes whose products are required for entry, proliferation, protection, and persistence in the preferred infection niches for extra- and intracellular pathogenic bacteria. Abundance of virulence factors are yet another response of pathogen for evading host immune system and adapting to host physiological conditions, e.g., pertussis toxin (PTX) secretion by *Bordetella pertussis* (*B. pertussis*) induces hyperinsulinemia and hypoglycemia conditions in infected host [50]. PTX is involved in catalysis of ADP-ribosylation of alpha subunits $G\alpha_{i/o}$ ($G_{i/o}$) protein family and interferes in intracellular signaling, and the increased insulin release causes hypoglycemia in host [50]. Microbial infections also alter the production of inflammatory cytokines such as tumor necrosis factor (TNF α levels), which activates the sympathetic nervous system and, thereby, reduces the blood glucose level (hypoglycemia), which is a significant observation as biomarker during severe malaria and cerebral malaria [51].

Amino acids are another metabolite product which is associated with interactions between host and pathogen and influences outcome of infection. At the infection site, both host and pathogen share similar nutritional substrates and generate common metabolic products; therefore, cross talk between their metabolic pathways could affect the pathogenesis of infection. The alteration in leucine/isoleucine, arginine, tryptophan, glutamine, and proline amino acids might influence the outcome of infection and regulation of host immune system [10, 11]. L-Glutamine induces expression of *Listeria monocytogenes* virulence gene factors [52]. *Vaccinia virus* alters the metabolic pathways of the host for efficient replication and fully relies on glutamine but not glucose to anaplerotically maintain TCA cycle (Fig. 22.3) [53]. During infection, the host imposes manganese and zinc starvation on invading pathogens, but altering cellular metabolism contributes to the ability of pathogens to resist manganese starvation and that ArlRS a global staphylococcal virulence regulator enables *Staphylococcus aureus* to overcome nutritional immunity by facilitating this adaptation. *S. aureus* adapts to the impaired glycolysis by switching its dependency for energy from glycolysis to amino acid metabolism pathway and reduces demands of manganese and zinc for their growth [54]. As a response of host cell to restrict *Leishmania* infection, L-arginine is metabolized through oxygenation by inducible nitric oxide synthase (iNOS) to form nitric oxide (NO) [55]. In order to inhibit this antimicrobial response of host, the pathogen increases the arginase activity, arginine transport, and arginine deiminase pathway to compete with iNOS for NO production and, thus, depletes the arginine [55]. Tryptophan is required for the proliferation of T cell and boost the immune system against microbial infection [56]. Indoleamine 2,3-dioxygenase (IDO) is a rate-limiting enzyme in the breakdown of the essential amino acid tryptophan to kynurenine. Interestingly, the microbial pathogen depletes the tryptophan pool from the cell by inducing the immunomodulator IDO to form kynurenine, which, in turn, promotes apoptosis of neutrophils and inhibits reactive oxygen species (ROS). Thus, the accumulation of kynurenine could serve as an indicator of *Mycobacterium* infection (*M. avium* subsp.

paratuberculosis and other virulent mycobacteria) [56]. The probiotics such as *Lactobacillus acidophilus* W22, *Bifidobacterium lactis* W51, *Lactobacillus brevis* W63, *Bifidobacterium bifidum* W23, *Lactococcus lactis* W58, and *Enterococcus faecium* W54 can be used as a supplement of tryptophan to increase its concentrations in serum and reduce the infections related to the upper respiratory tract [10, 11]. Furthermore, elevated amino acid catabolism and folate and lipoic acid biosynthesis were observed in sputum of cystic fibrosis patients, which indicated the evolution of microbial pathogens against sulfonamide drugs which target folate biosynthesis [57]. Increased level of amino acids such as histidine, phenylalanine, glutamine, methionine, lysine, leucine, tryptophan, tyrosine, and glutamic acid were observed during white spot syndrome virus infection to evade the host immune system and facilitate increased viral replication and multiplication [58]. Microbial infections especially virus invasions in host induce synthesis of nucleotide, protein, fatty acid, and cholesterol for viral replication, and the virus remodels the host intracellular membrane for assembly of virus with host cell components [59, 60]. Elevated levels of fatty acid synthesis and glutaminolysis are found in cells infected with viruses like human cytomegalovirus, influenza A, hepatitis C, etc. and such modifications of carbon source utilization facilitate virus replication and virion production by increasing energy available (Fig. 22.3) [61]. In response, host immune system activates 5'-AMP-activated protein kinase (AMPK) to inhibit the fatty acid synthesis [62]. However, AMPK signaling could facilitate or inhibit intracellular viral replication depending on the microbial infection [62]. Alterations in lipid metabolism were found to be triggered by Zika virus infection, and infected patients showed increased levels of several phosphatidylethanolamine (PE) lipid species in serum, especially plasmeyl-phosphatidylethanolamine (pPE) (or plasmalogens) linked to polyunsaturated fatty acids, indicating increased viral replication and infectivity [63]. Biosynthesis of plasmalogens requires functional peroxisomes, which are important sites for viral replication. Influenza A virus promotes synthesis of prostaglandin E2 (PGE₂), an eicosanoid generated by cyclooxygenases, and inhibits recruitment and activity of macrophages; thus the virus suppresses both innate and adaptive immunity [64]. Analysis of COVID-19 patient samples exhibited plasma metabolome changes, indicating perturbed oxidative pathways of cellular energy production [65]. The levels of acylcarnitines (palmitoylcarnitine, oleoylcarnitine, and stearoylcarnitine) and metabolites associated with tricarboxylic acid (TCA) cycle were reduced, indicating attenuated entry of fatty acyls into the mitochondria for β -oxidation. Lactate dehydrogenase level was increased with disease severity, whereas significantly reduced level of numerous amino acid (tryptophan, proline, valine, isoleucine, and citrulline) was found in mild and moderate COVID patients [65].

Metabolomics provides the comprehensive information on metabolites altered during infection and offers scope to explore through metabolite profiling the dynamics of adaptations, which contributed to the microbial death or survival to become resistant. Metabolite profiling of ampicillin-resistant *E. coli* revealed that recycling of anhydro-muropeptides plays an important role in mediating resistance to ampicillin [66, 67]. *E. coli* breaks down over 60% of the murein of its side wall and reuses the component amino acids to synthesize about 25% of the cell wall for the next

generation, and 1,6-anhydro-*N*-acetylmuramic acid (anhMurNAc) is returned to the biosynthetic pathway by conversion to *N*-acetylglucosamine-phosphate (GlcNAc-P) [66, 67]. A study reported the accumulation of dTDP-rhamnose (a deoxy sugar) in *E. coli* cells treated with quinolones (nalidixic acid and norfloxacin) and that increasing concentrations of dTDP-rhamnose upregulated the gyrase A transcription and helped the cells cope with antibiotic by sequestering the antibiotic and reducing drug-gyrase complex formation [68]. *N*-Acetyl-glutamate was accumulated in *E. coli* in response to kanamycin and spectinomycin antibiotics, which might be a hallmark of protein synthesis inhibition [68]. Multidrug-resistant (MDR) strains of *E. coli* with resistance to ciprofloxacin, ampicillin, gentamicin, and sulfamethoxazole showed increased levels of L-serine, glycine including L-cystine, L-cysteine-sulfinate, S-sulfo-L-cysteine, L-cystathionine, L-methionine sulfoxide, and L-methionine metabolites [69]. Altered pathways for amino acids, phenylpropanoids, and purine metabolism resulted in altered glucose, fatty acids, and ammonia biosynthesis which might help in acquiring drug resistance and survival under stress conditions [69]. Upregulation of cysteine and methionine amino acids involves redox reactions in bacteria, which prevent the entry of drugs into the cells and result in drug resistance [69]. MDR (fluconazole, ketoconazole, and miconazole) strains of *C. albicans* exhibited increased drug efflux; higher membrane fluidity; significantly reduced levels of N-methylnicotinate, glycerol, N-dimethylglycine, ribitol, proline, L-aspartate, valine, glutamine, and N-acetyl-aspartic acid metabolites; but increased levels of glycerophosphocholine, L-aspartate-4-semialdehyde, 2-oxoglutarate, adenine, lysophosphatidylcholine (16:1), phytosphingosine-1-P, Cer(d18:0/14:0), serine, spermine, lactate, inosine, citrate, 2-deoxyribose, and succinate, predicting high tolerance of *C. albicans* cells against drugs, oxidative stress, and temperature variations [70].

Recently, using untargeted metabolomics, Diaz et al. studied differential metabolites between *M. tuberculosis* and live attenuated *M. tuberculosis* vaccine named MTBVAC and correlated the vaccine with its parental strain based on metabolite profiling [71]. Identification of some of the differential metabolites might be useful as potential vaccination biomarkers [71]. Studies clearly indicate that carbohydrate, amino acid, and fatty acid metabolism pathways are majorly affected during microbial infection, adaptation, and evolution. The altered metabolites of these cellular processes accumulate in host cell and may serve as a target for controlling disease severity and progression. Metabolite profiling certainly advances our understandings on host-microbial interactions, metabolic pathways, and targets which can be used for the development of effective therapeutics and treatments.

22.4 Conclusion

Microbial infections have become impossible to eradicate, and they are still causing epidemics and pandemics and affecting millions of lives. Metabolomics has emerged as a powerful approach, which directly represents the molecular phenotype and reflects the underlying cellular processes and alterations in infected or diseased

state. Metabolomics provides the information which can be translated into assays or technology for easy and rapid disease diagnosis, identification of microbial strains, and study of host-pathogen interactions to aid development of effective therapeutics and disease treatment.

22.5 Future Perspectives

High-sensitivity and high-throughput metabolomics have enabled comprehensive detection of thousands of small-molecule metabolites in host and microbial communities. In the last 20 years, advances in techniques for metabolomics, availability of databases for metabolites, and analytical software have contributed to establishing new methods for disease diagnosis, metabolic studies, improved treatment, and molecular phenotyping of cellular processes. Recent technique such as high-resolution metabolomics can simultaneously identify the metabolic pathways and associated inflammatory cytokines, which can save the time for the elucidation of host immune response upon microbial invasion. Importantly, metabolomics has allowed clinicians to ameliorate potential effects of the errors occurred during metabolism and adaptation of pathogen in host. Metabolomics data can be utilized by computational biologists to design models for microbial pathogenesis and to identify new drug targets, especially against MDR pathogens, so that personalized therapy can be developed (Fig. 22.4). Metabolomics offers new opportunities for biomarker discovery in complex diseases and may provide pathological understanding of diseases beyond traditional technologies.

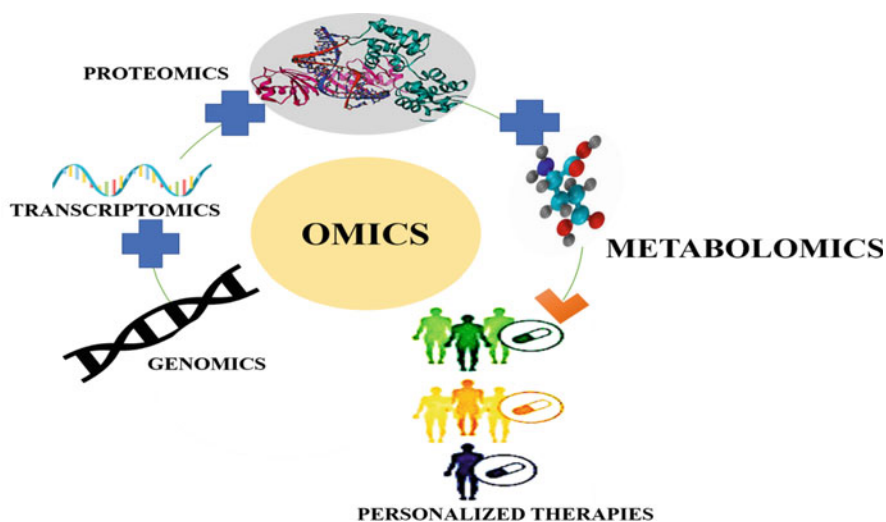
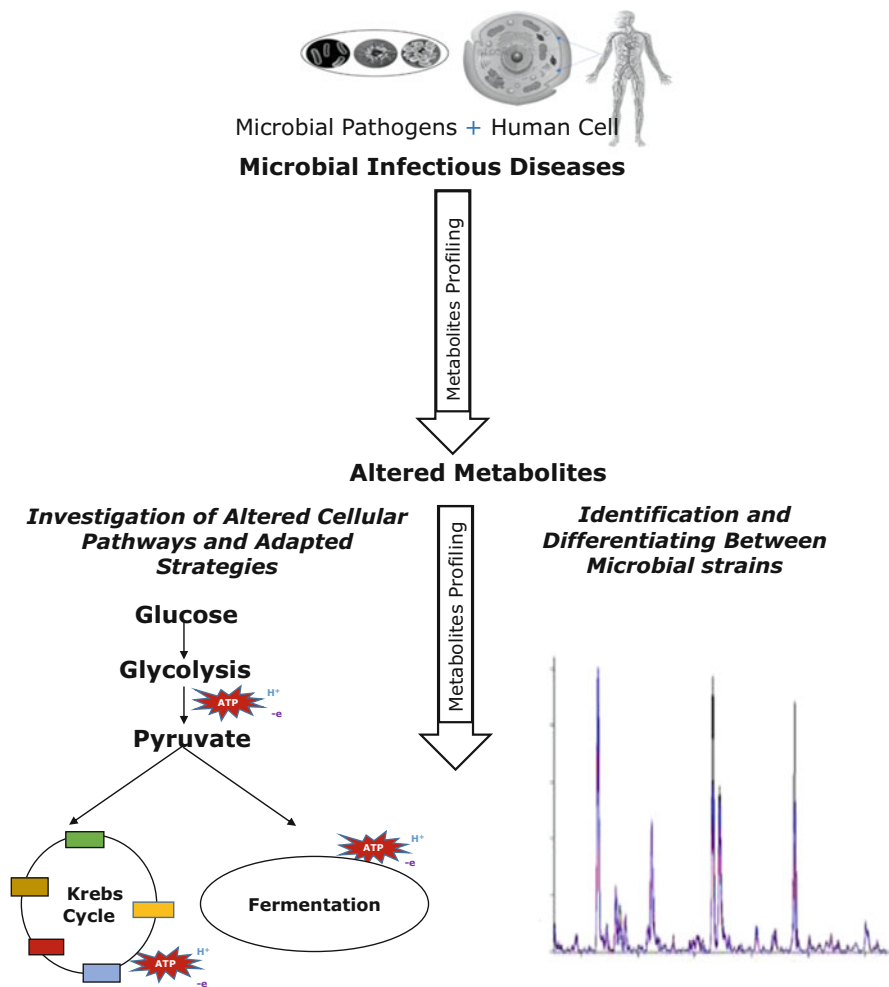


Fig. 22.4 Future perspectives of functional metabolomics



Scheme 22.1 Graphical representation of abstract

Acknowledgments TP acknowledges JNU-UPOE II scheme [ID-161] and JNU-DST-PURSE (Phase-II). Preeti acknowledges the Council of Scientific and Industrial Research (CSIR), India, for providing Senior Research Fellowship, and Megha acknowledges Jawaharlal Nehru University (JNU) for fellowship.

References

1. Clish CB (2015) Metabolomics: an emerging but powerful tool for precision medicine. Cold Spring Harb Mol Case Stud 1(1):a000588. <https://doi.org/10.1101/mcs.a000588>

2. Kim SJ, Kim SH, Kim JH et al (2016) Understanding metabolomics in biomedical research. *Endocrinol Metab (Seoul)* 31(1):7–16. <https://doi.org/10.3803/EnM.2016.31.1.7>
3. Shen B, Yi X, Sun Y et al (2020) Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell* 182(1):59–72.e15. <https://doi.org/10.1016/j.cell.2020.05.032>
4. Fernández-García M, Rojo D, Rey-Stolle F et al (2018) Metabolomic-based methods in diagnosis and monitoring infection progression. In: Silvestre R, Torrado E (eds) *Metabolic interaction in infection. Experientia Supplementum*, vol 109. Springer, Cham. https://doi.org/10.1007/978-3-319-74932-7_7
5. Marshall DD, Powers R (2017) Beyond the paradigm: combining mass spectrometry and nuclear magnetic resonance for metabolomics. *Prog Nucl Magn Reson Spectrosc* 100:1–6. <https://doi.org/10.1016/j.pnmrs.2017.01.001>
6. Mesías AC, Garg NJ, Zago MP (2019) Redox balance keepers and possible cell functions managed by redox homeostasis in *Trypanosoma cruzi*. *Front Cell Infect Microbiol* 9:435. <https://doi.org/10.3389/fcimb.2019.00435>
7. Hertel M, Hartwig S, Schütte E et al (2016) Identification of signature volatiles to discriminate *Candida albicans*, *glabrata*, *krusei* and *tropicalis* using gas chromatography and mass spectrometry. *Mycoses* 59(2):117–126. <https://doi.org/10.1111/myc.12442>
8. Rahi P, Vaishampayan P (2020) MALDI-TOF MS applications in microbial ecology studies. *Front Microbiol* 10:2954. <https://doi.org/10.3389/fmicb.2019.02954>
9. Introduction-of-NIST-17 (n.d.). <https://www.americanlaboratory.com/913-Technical-Articles/340911-Introduction-of-NIST-17-A-Major-Update-of-Mass-Spectral-Libraries-and-Software-at-the-65th-ASMS-Conference-on-Mass-Spectrometry-and-Allied-Topics>
10. Ren JL, Zhang AH, Kong L et al (2018a) Advances in mass spectrometry-based metabolomics for investigation of metabolites. *RSC Adv* 8(40):22335–22350. <https://doi.org/10.1039/c8ra01574k>
11. Ren W, Rajendran R, Zhao Y et al (2018b) Amino acids as mediators of metabolic cross talk between host and pathogen. *Front Immunol* 319(018):9. <https://doi.org/10.3389/fimmu.2018.00319>
12. Marko DC, Saffert RT, Cunningham SA et al (2012) Evaluation of the Bruker Biotyper and Vitek MS matrix-assisted laser desorption ionization–time of flight mass spectrometry systems for identification of nonfermenting gram-negative bacilli isolated from cultures from cystic fibrosis patients. *Eur J Clin Microbiol* 50(6):2034–2039. <https://doi.org/10.1128/JCM.00330-12>
13. Hou L, Wei X, Zhuo Y et al (2018) GC-MS-based metabolomics approach to diagnose depression in hepatitis B virus-infected patients with middle or old age. *Aging (Albany NY)* 10(9):2252. <https://doi.org/10.18632/aging.101535>
14. Miura D, Fujimura Y, Yamato M et al (2010) Ultrahighly sensitive in situ metabolomic imaging for visualizing spatiotemporal metabolic behaviors. *Anal Chem* 82(23):9789–9796. <https://doi.org/10.1021/ac101998z>
15. Emwas AH, Roy R, McKay RT et al (2019) NMR spectroscopy for metabolomics research. *Meta* 9(7):123. <https://doi.org/10.3390/metabo9070123>
16. Dona AC, Kyriakides M, Scott F et al (2016) A guide to the identification of metabolites in NMR-based metabolomics/metabolomics experiments. *Comput Struct Biotechnol* 14:135–153. <https://doi.org/10.1016/j.csbj.2016.02.005>
17. Jeong S, Eskandari R, Park SM et al (2017) Real-time quantitative analysis of metabolic flux in live cells using a hyperpolarized micromagnetic resonance spectrometer. *Sci Adv* 3:e1700341. <https://doi.org/10.1126/sciadv.1700341>
18. Motta A, Paris D, Melck D (2010) Monitoring real-time metabolism of living cells by fast two-dimensional NMR spectroscopy. *Anal Chem* 82(6):2405–2411. <https://doi.org/10.1021/ac9026934>
19. Milner JJ, Wang J, Sheridan PA et al (2014) ¹H NMR-based profiling reveals differential immune-metabolic networks during influenza virus infection in obese mice. *PLoS One* 9(5):97238. <https://doi.org/10.1371/journal.pone.0097238>

20. Lussu M, Camboni T, Piras C et al (2017) H NMR spectroscopy-based metabolomics analysis for the diagnosis of symptomatic *E. coli*-associated urinary tract infection (UTI). *BMC Microbiol* 17(1):201. <https://doi.org/10.1186/s12866-017-1108-1>
21. Ntziachristos V, Pleitez MA, Aime S et al (2019) Emerging technologies to image tissue metabolism. *Cell Metab* 29(3):518–538. <https://doi.org/10.1016/j.cmet.2018.09.004>
22. Freedberg DI, Selenko P (2014) Live cell NMR. *Annu Rev Biophys* 43:171. <https://doi.org/10.1146/annurev-biophys-051013-023136>
23. Tsedilin AM, Fakhrutdinov AN, Eremin DB et al (2015) How sensitive and accurate are routine NMR and MS measurements. *Mendelev Commun* 25(6):454. <https://doi.org/10.1016/j.mencom.2015.11.019>
24. Giraudeau P (2020) NMR-based metabolomics and fluxomics: developments and future prospects. *Analyst* 145(7):2457–2472. <https://doi.org/10.1039/D0AN00142B>
25. Judge MT, Wu Y, Tayyari F et al (2019) Continuous in vivo metabolism by NMR. *Front Mol Biosci* 6:26. <https://doi.org/10.1128/JVI.03134-13>
26. Zhu J, Bean HD, Kuo YM et al (2010) Fast detection of volatile organic compounds from bacterial cultures by secondary electrospray ionization-mass spectrometry. *J Clin Microbiol* 49(2):769. <https://doi.org/10.1128/JCM.00392-10>
27. Geldreich EE, Kenner BA, Kabler PW (1964) Occurrence of coliforms, fecal coliforms, and *Streptococci* on vegetation and insects. *Appl Microbiol* 12(1):63–69. <https://doi.org/10.1128/aem.12.1.63-69.1964>
28. Phillips M, Basa-Dalay V, Blais J et al (2012) Point-of-care breath test for biomarkers of active pulmonary tuberculosis. *Tuberculosis* 92(4):314–320. <https://doi.org/10.1016/j.tube.2012.04.002>
29. Enderby B, Smith D, Carroll W et al (2009) Hydrogen cyanide as a biomarker for *Pseudomonas aeruginosa* in the breath of children with cystic fibrosis. *Pediatr Pulmonol* 44(2):142–147. <https://doi.org/10.1002/ppul.20963>
30. Scott-Thomas AJ, Syhre M, Pattermore PK et al (2010) 2-Aminoacetophenone as a potential breath biomarker for *Pseudomonas aeruginosa* in the cystic fibrosis lung. *BMC Pulm Med* 10:56. <https://doi.org/10.1186/1471-2466-10-56>
31. Tong H, Wang Y, Li Y et al (2017) Volatile organic metabolites identify patients with gastric carcinoma, gastric ulcer, or gastritis and control patients. *Cancer Cell Int* 17(1):1–9. <https://doi.org/10.1186/s12935-017-0475-x>
32. Garner CE, Smith S, Bardhan PK et al (2009) A pilot study of faecal volatile organic compounds in faeces from cholera patients in Bangladesh to determine their utility in disease diagnosis. *Trans R Soc Trop Med Hyg* 103(11):1171–1173. <https://doi.org/10.1016/j.trstmh.2009.02.004>
33. Koo S, Thomas HR, Daniels SD et al (2014) A breath fungal secondary metabolite signature to diagnose invasive aspergillosis. *Clin Infect Dis* 59(12):1733–1740. <https://doi.org/10.1093/cid/ciu725>
34. Briard B, Heddergott C, Latg JP (2016) Volatile compounds emitted by *Pseudomonas aeruginosa* stimulate growth of the fungal pathogen *Aspergillus fumigatus*. *MBio* 7(2). <https://doi.org/10.1128/mBio.00219-16>
35. Wikoff WR, Anfora AT, Liu J et al (2009) Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc Natl Acad Sci U S A* 106(10):3698–3703. <https://doi.org/10.1073/pnas.0812874106>
36. Pabary R, Huang J, Kumar S et al (2016) Does mass spectrometric breath analysis detect *Pseudomonas aeruginosa* in cystic fibrosis? *Eur Respir J* 47(3):994–997. <https://doi.org/10.1183/13993003.00944-2015>
37. Miekisch W, Schubert JK, Noeldge-Schomburg GF (2004) Diagnostic potential of breath analysis—focus on volatile organic compounds. *Clin Chim Acta* 347(1–2):25–39. <https://doi.org/10.1016/j.cccn.2004.04.023>
38. Carroll W, Lenney W, Wang T et al (2005) Detection of volatile compounds emitted by *Pseudomonas aeruginosa* using selected ion flow tube mass spectrometry. *Pediatr Pulmonol* 39(5):452–456. <https://doi.org/10.1002/ppul.20170>

39. Gilchrist FJ, Belcher J, Jones AM et al (2015) Exhaled breath hydrogen cyanide as a marker of early *Pseudomonas aeruginosa* infection in children with cystic fibrosis. *ERJ Open Res* 1:00044–02015. <https://doi.org/10.1183/23120541.00044-2015>
40. Robroeks CM, van Berkel JJ, Dallinga JW et al (2010) Metabolomics of volatile organic compounds in cystic fibrosis patients and controls. *Pediatric Res* 68(1):75–80. <https://doi.org/10.1203/PDR.0b013e3181df4ea0>
41. Sethi S, Nanda R, Chakraborty T (2013) Clinical application of volatile organic compound analysis for detecting infectious diseases. *Clin Microbiol Rev* 26(3):462–475. <https://doi.org/10.1128/CMR.00020-13>
42. Traxler S, Barkowsky G, Saß R et al (2019) Volatile scents of *influenza A* and *S. pyogenes* (co-) infected cells. *Sci Rep* 9(1):18894. <https://doi.org/10.1038/s41598-019-55334-0>
43. Fend R, Kolk AH, Bessant C et al (2006) Prospects for clinical application of electronic-nose technology to early detection of *Mycobacterium tuberculosis* in culture and sputum. *J Clin Microbiol* 44(6):2039–2045. <https://doi.org/10.1128/JCM.01591-05>
44. Ahmed WM, Lawal O, Nijssen TM et al (2017) Exhaled volatile organic compounds of infection: a systematic review. *ACS Infect Dis* 3(10):695–710. <https://doi.org/10.1021/acsinfectdis.7b00088>
45. Amarante-Mendes GP, Adjemian S, Branco LM (2018) Pattern recognition receptors and the host cell death molecular machinery. *Front Immunol* 9:2379. <https://doi.org/10.3389/fimmu.2018.02379>
46. Eisenreich W, Rudel T, Heesemann J et al (2019) How viral and intracellular bacterial pathogens reprogram the metabolism of host cells to allow their intracellular replication. *Front Cell Infect Microbiol* 9:42. <https://doi.org/10.3389/fcimb.2019.00042>
47. Barisón MJ, Rapado LN, Merino EF et al (2017) Metabolomic profiling reveals a finely tuned, starvation-induced metabolic switch in *Trypanosoma cruzi* pimatigotes. *JBiolChem* 292(21):8964–8977. <https://doi.org/10.1074/jbc.M117.778522>
48. Cumming BM, Pacl HT, Steyn AJ (2020) Relevance of the Warburg effect in tuberculosis for host-directed therapy. *Front Cell Infect Microbiol* 10:506. <https://doi.org/10.3389/fcimb.2020.576596>
49. Fuchs TM, Eisenreich W, Heesemann J et al (2012) Metabolic adaptation of human pathogenic and related non-pathogenic bacteria to extra-and intracellular habitats. *FEMS Microbiol Rev* 36(2):435–462. <https://doi.org/10.1111/j.1574-6976.2011.00301.x>
50. Melvin JA, Scheller EV, Miller JF et al (2014) *Bordetella pertussis* pathogenesis: current and future challenges. *Nat Rev Microbiol* 12(4):274–288. <https://doi.org/10.1038/nrmicro3235>
51. Mavondo GA, Mavondo J, Peresuh W et al (2019) Malaria Pathophysiology as a Syndrome: Focus on glucose homeostasis in severe malaria and phytotherapeutics management of the Disease. In *Parasites and Parasitic Diseases*. Intech Open. <https://doi.org/10.5772/intechopen.79698>
52. Haber A, Friedman S, Lobel L et al (2017) L-glutamine induces expression of *Listeria monocytogenes* virulence genes. *PLoS Pathog* 13(1):e1006161. <https://doi.org/10.1371/journal.ppat.1006161>
53. Fontaine KA, Camarda R, Lagunoff M (2014) Vaccinia virus requires glutamine but not glucose for efficient replication. *J Virol* 88(8):4366–4374. <https://doi.org/10.1128/JVI.03134-13>
54. Radin JN, Kelliher JL, Parraga Solorzano PK et al (2016) The two-component system ArIRS and alterations in metabolism enable *Staphylococcus aureus* to resist calprotectin-induced manganese starvation. *PLoS Pathog* 12(11):e1006040. <https://doi.org/10.1371/journal.ppat.1006040>
55. Olekhovitch R, Bouso P (2015) Induction, propagation, and activity of host nitric oxide: lessons from *Leishmania* infection. *Trends Parasitol* 31(12):653–664. <https://doi.org/10.1016/j.pt.2015.08.001>

56. Plain KM, de Silva K, Earl J et al (2011) Indoleamine 2,3-dioxygenase, tryptophan catabolism, and *Mycobacterium avium* subsp. *paratuberculosis*: a model for chronic mycobacterial infections. *Infect Immun* 79(9):3821–3832. <https://doi.org/10.1128/IAI.05204-11>
57. Quinn RA, Lim YW, Maughan H et al (2014) Biogeochemical forces shape the composition and physiology of polymicrobial communities in the cystic fibrosis lung. *MBio* 5(2):e00956–e00913. <https://doi.org/10.1128/mBio.00956-13>
58. Fan W, Ye Y, Chen Z et al (2016) Metabolic product response profiles of *Cherax quadricarinatus* towards white spot syndrome virus infection. *Dev Comp Immunol* 61:236–241. <https://doi.org/10.1016/j.dci.2016.04.006>
59. Heaton NS, Randall G (2011) Multifaceted roles for lipids in viral infection. *Trends Microbiol* 19:368–375. <https://doi.org/10.1016/j.tim.2011.03.007>
60. Miller S, Krijnse-Locker J (2008) Modification of intracellular membrane structures for virus replication. *Nat Rev Microbiol* 6:363–374. <https://doi.org/10.1038/nrmicro1890>
61. Sanchez EL, Lagunoff M (2015) Viral activation of cellular metabolism. *Virology* 479–480:609–618. <https://doi.org/10.1016/j.virol.2015.02.038>
62. Silwal P, Kim JK, Yuk JM et al (2018) AMP-activated protein kinase and host defense against infection. *Int J Mol Sci* 19(11):3495. <https://doi.org/10.3390/ijms19113495>
63. Queiroz A, Pinto IFD, Lima M et al (2019) Lipidomic analysis reveals serum alteration of plasmalogens in patients infected with ZIKA virus. *Front Microbiol* 10:753. <https://doi.org/10.3389/fmicb.2019.00753>
64. Sander WJ, O'Neil HG, Pohl CH et al (2017) Prostaglandin E₂ as a modulator of viral infections. *Front Physiol* 8:89. <https://doi.org/10.3389/fphys.2017.00089>
65. Song JW, Lam SM, Fan X et al (2020) Omics-driven systems interrogation of metabolic dysregulation in COVID-19 pathogenesis. *Cell Metab* 32(2):188–202. <https://doi.org/10.1016/j.cmet.2020.06.016>
66. Johnson JW, Fisher JF, Mobashery S (2013) Bacterial cell-wall recycling. *Ann N Y Acad Sci* 1277(1):54. <https://doi.org/10.1111/j.1749-6632.2012.06813.x>
67. Zampieri M, Enke T, Chubukov V et al (2017a) Metabolic constraints on the evolution of antibiotic resistance. *Mol Syst Biol* 13(3):917. <https://doi.org/10.15252/msb.20167028>
68. Zampieri M, Zimmermann MC et al (2017b) Nontargeted metabolomics reveals the multilevel response to antibiotic perturbations. *Cell Rep* 19(6):1214–1228. <https://doi.org/10.1016/j.celrep.2017.04.002>
69. Lin Y, Li W, Sun L et al (2019) Comparative metabolomics shows the metabolic profiles fluctuate in multi-drug resistant *Escherichia coli* strains. *J Proteome* 207:103468. <https://doi.org/10.1016/j.jprot.2019.103468>
70. Li L, Liao Z, Yang Y et al (2018) Metabolomic profiling for the identification of potential biomarkers involved in a laboratory azole resistance in *Candida albicans*. *PLoS One* 13(2):e0192328. <https://doi.org/10.1371/journal.pone.0192328>
71. Díaz C, del Palacio JP, Valero-Guillén PL et al (2019) Comparative metabolomics between *Mycobacterium tuberculosis* and the MTBVAC vaccine candidate. *ACS Infectious Diseases* 5(8):1317–1326. <https://doi.org/10.1021/acsinfectdis.9b00008>



Lipidomics to Study the Role of Lipid Droplets in Host-Pathogen Interactions

23

Anwesha Bhattacharyya and Vineet Choudhary

Abstract

Lipid droplets (LDs) are intracellular organelles dedicated for fat storage and play critical roles in cellular homeostasis. Recently, LD biology has moved to the forefront of biomedical research due to their involvement in a variety of diseases that are affected by lipid imbalance, such as obesity, type 2 diabetes, fatty liver, cardiovascular diseases, Alzheimer's disease, and cancer. Growing evidence suggests that majority of intracellular pathogens, be they viral, bacterial, or protozoan, rely on host LDs for completing some steps of their life cycle, thus emphasizing the importance of LDs in host-pathogen interactions. Host and pathogen lipids play vital role in the ability of the pathogen to evade host immune system. Therefore, droplet homeostasis and pathogen replication are intricately linked, the mechanisms of which are largely unknown. This chapter summarizes our current understanding of how unique aspect of LD biology is exploited by pathogens for their replication and propagation in the host. Advancement in the field of lipidomics for performing lipid-profiling of host-pathogen interactions will shed light on many novel and unanticipated findings in disease pathogenesis aimed at discovery of novel biomarkers and identification of therapeutic interventions.

A. Bhattacharyya

Multidisciplinary Centre for Advance Research and Studies (MCARS), Jamia Millia Islamia, Jamia Nagar, Okhla, New Delhi, India

V. Choudhary (✉)

Department of Biotechnology, All India Institute of Medical Sciences (AIIMS), New Delhi, India
e-mail: vchoudhary@aiims.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

425

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*,
https://doi.org/10.1007/978-981-16-0691-5_23

Keywords

Lipid droplets · LDs · Intracellular pathogen · Host-pathogen interactions · Lipoprotein biogenesis · HCV · *Dengue virus* · *Rotavirus* · *Chlamydia* · *Mycobacteria* · Parasite

23.1 Introduction

Lipid droplets (LDs) are cytoplasmic fat storage organelles found in all cell types. The core of LDs comprises neutral lipids (NLs), triacylglycerols (TG), and sterol esters (SE) surrounded by a monolayer of phospholipids that harbors growing number of associated proteins [1]. LDs originate in the endoplasmic reticulum (ER), where NL-synthesizing proteins reside. DGAT1 and DGAT2 (diacylglycerol acyltransferases, Dga1, and Lro1 in yeast) catalyze the TG synthesis, whereas ACAT1 and ACAT2 (acyl-coenzyme A:cholesterol acyltransferases, ARE1, and ARE2 in yeast) produce SE. NLs accumulate inside the bilayer membrane of the ER, which upon reaching a threshold gets oiled out between the ER bilayer membrane resulting in lens-like structure formation. These lipid lenses grow in size by acquiring more NLs and eventually emerge toward the cytoplasm where they further mature [1, 2] (Fig. 23.1). Recent studies have shed light on the mechanistic process of LD emergence, revealing that this process is mediated by biophysical property of locally enriched lipids, and LD assembly factors that assemble at specialized tubular ER subdomains [2–6]. Local changes in asymmetry of lipids, and packing at LD biogenesis sites, alter surface tension, membrane curvature properties, and lateral pressure that affect the directionality of LD budding [6]. Very recently, proteins that define LD biogenesis sites have been identified. Seipin together with lipid droplet assembly factor 1 (LDAF1) were found to mark discrete ER sites of LD biogenesis in mammalian cells [7]. Similarly yeast seipin (Fld1/Sei1) and Nem1 (catalytic subunit of Nem1-Spo7 phosphatase complex that regulates diacylglycerol production) were found to localize to discrete ER subdomains where they both together (Fld1-Nem1) play a crucial role in initiating LD biogenesis [4].

LD surface monolayer is decorated by many lipid-modifying enzymes, including acyltransferases, and lipases, together with structural proteins such as perilipin (Plin) family members (Pet10 in yeast). Plin1 and Plin2 are key LD coat proteins that regulate biogenesis and degradation of LDs [1]. Upon lipolytic stimulation, adipose triglyceride lipase (ATGL) degrades TG to diacylglycerol (DG), which is subsequently cleaved by hormone-sensitive lipase (HSL) to monoacylglycerol (MG) [8, 9]. Enzyme phospholipase A (PLA) mediates the breakdown of arachidonic acids (AA) from LDs, thereby producing eicosanoid immune modulators, including leukotrienes and prostaglandins [10]. LDs have multitude of functions apart from roles in lipid metabolism, including membrane trafficking and signal transduction [11], ER stress response [12], protection from lipotoxicity, protein degradation, and regulation of autophagy [1]. LDs are implicated in many pathological conditions such as type 2 diabetes, obesity, atherosclerosis, fatty liver,

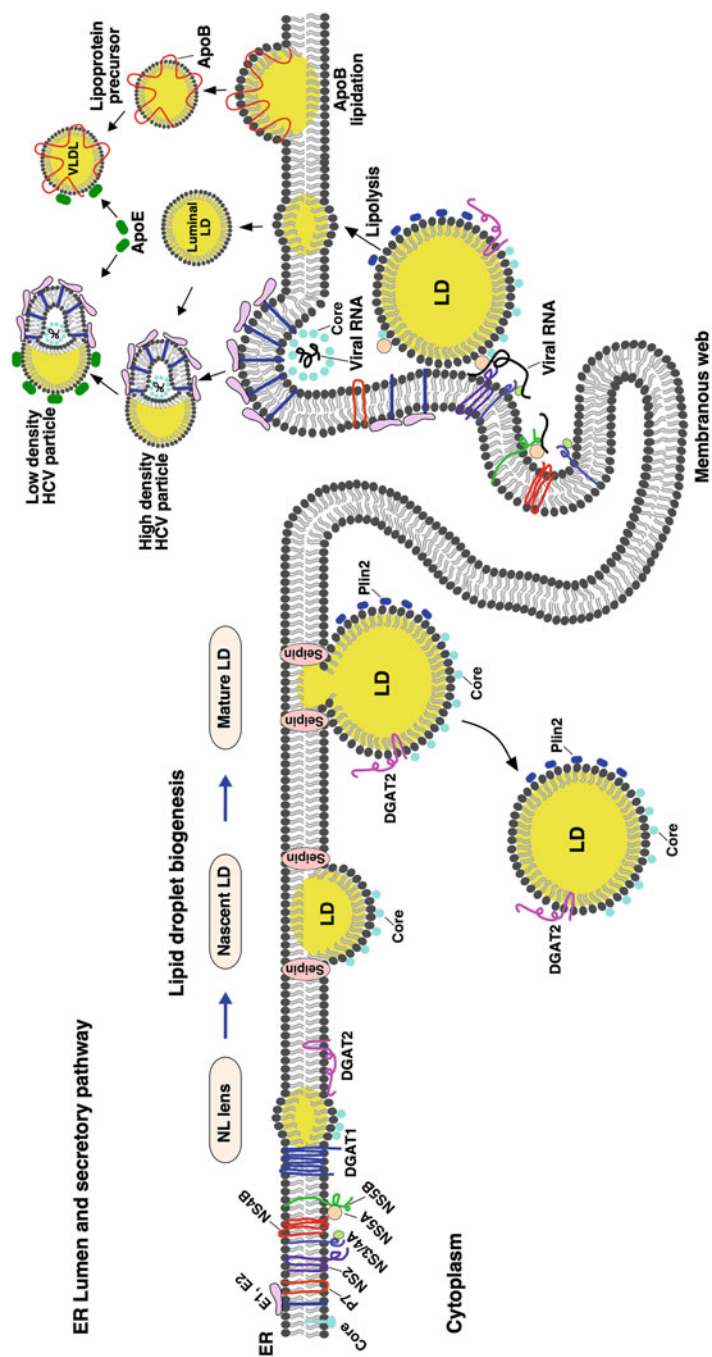


Fig. 23.1 Model of lipid droplet biogenesis and HCV assembly. Neutral lipids (NL) are synthesized by ER-localized triacylglycerol-producing enzymes (DGAT, diglyceride acyltransferase) and sterol ester-synthesizing enzymes (ACAT, acyl-coenzyme A:cholesterol acyltransferase), resulting in formation of NL lenses inside the ER bilayer membrane that grows into nascent LDs at specialized ER subdomains defined by seipin protein. Nascent LDs emerge toward the cytoplasm where they further mature. LD coat proteins, such as Plin2, decorate the monolayer surface of LDs. Mature LDs may remain associated with the ER or might detach from the ER as an independent organelle in the cytoplasm. During HCV life cycle, the viral RNA is translated into one polyprotein at the ER, cleaved by host and viral proteases to release ten viral proteins. The viral nucleocapsid core protein undergoes two subsequent cleavages, thereby releasing the

cancer, and some neurodegenerative diseases [1]. Unique properties of LDs make them an attractive target for several pathogens to interact with LDs for completing some steps of their life cycle. Thus, hepatitis C virus (HCV), dengue virus, rotavirus, a number of intracellular bacteria, and parasites hijack host LDs for their multiplication (reviewed in [13]). In this chapter we will discuss the role of LDs in host-pathogen interactions that is crucial for the replication and propagation of pathogens inside host.

23.2 Emerging Role of Lipid Droplets in Pathogen Replication Cycle

23.2.1 LDs and Viral Infection

Hepatitis C Virus—HCV is the best characterized human pathogen having close interactions with host LDs. HCV is a blood-borne virus that belongs to the *Flaviviridae* family and mostly replicates in hepatocytes. The HCV is internalized by receptor-mediated endocytosis utilizing a combination of lipid and lipoprotein receptors that are unique to the hepatocytes (e.g., scavenger receptor B type 1 (SR-BI), low-density lipoprotein receptor (LDL-R), claudin-1, and occludin) [14]. Upon uncoating and release, the positive-stranded viral RNA is translated at the ER which is cleaved by host and viral proteases releasing three proteins: three structural proteins (the nucleocapsid core and the two envelope glycoproteins E1 and E2), the p7 viroporin, and six nonstructural proteins (NS2, NS3, NS4A, NS4B, NS5A, and NS5B) [15, 16]. RNA replication complexes consisting of NS3-NS5B proteins replicate the viral RNA within the ER-derived structures, the so-called membranous web (Fig. 23.1). Interestingly, cytoplasmic LDs serve as platforms for the assembly of HCV virions, facilitating the translocation of the nucleocapsid core protein onto LD periphery in a DGAT1-dependent manner, thereby recruiting viral RNA replication machinery adjacent to the ER membrane in vicinity of LDs for encapsidation of newly synthesized viral RNA [17, 18]. TG synthesis activity of DGAT1 is essential for the core to access LDs, thereby assisting in successful HCV particle assembly. Thus, pharmacological inhibition of DGAT1 activity by using C75, an inhibitor of the fatty acid synthase enzyme complex (FASN), impairs HCV

Fig. 23.1 (continued) core that traffics onto LD monolayer in a DGAT1-dependent manner. The nonstructural proteins (NS3-NS5B) form RNA replication complexes in ER-derived membranous structure called “membranous web” in the vicinity of core-loaded LDs. NS3/NS4A and NS5A facilitate the transfer of newly synthesized viral genomes to the assembly sites. P7 and NS2 facilitate connections between core proteins and replication complex to the glycoproteins E1 and E2. Upon encapsidation of the viral RNA, viroparticles emerge toward the ER lumen having ER membrane surrounding it. Viroparticles interact with the host lipoprotein machinery and fuses to form lipo-viroparticles (high-/low-density HCV particle) that are secreted out of the cell along the secretory pathway

genome replication and particle maturation [19]. Membrane-spanning N-terminal region of the core protein is cleaved by two subsequent cleavage events by the action of signal peptidase (SP) and signal peptide peptidase (SPP), thereby liberating mature protein that traffics along the ER onto the droplet surface. The hydrophobic domain (amino acids 119–179) of the core protein comprises amphipathic helix that mediates tight association of the core with the LD monolayer and hence perturbs targeting of normal LD-resident proteins, such as Plin2, leading to microtubule-dependent clustering of core-coated LDs around nucleus [20]. HCV mutants, in which the core retains the C-terminal membrane-spanning domain, restrict the core to the ER and prevent its association with LDs, therefore producing less infectious virions, hence confirming that LDs play critical role in HCV replication [18]. Newly made viral RNA is encapsidated, and the resulting viroparticles emerge toward the ER lumen, interact with lipoprotein particles being assembled in the ER luminal compartment, and mature as lipo-viroparticles that subsequently exit the cell via classical secretory pathway [21] (Fig. 23.1).

The lipidome of HCV is similar to that of very low-density lipoproteins (VLDL) and low-density lipoproteins (LDL) and dissimilar to that of host cell organelles. During virion assembly, along with three structural proteins, various apolipoproteins are incorporated in the envelope including ApoB and the exchangeable apolipoproteins ApoA-I, ApoC-I, ApoC-II, and Apo-E that facilitate entry of virions inside host cells and assist in evading host immune response [22]. Taken together, maturation of HCV is tightly connected with the biogenesis of LDs and lipoproteins [16].

Dengue Virus—Dengue is caused by the dengue virus (DENV) that belongs to the mosquito-borne (*Aedes aegypti*) *Flavivirus*. Similar to HCV, DENV is a single-copy, positive-stranded RNA virus that belongs to the *Flaviviridae* family. DENV affects millions of people worldwide, mostly in warmer climates, where the mosquito thrives and causes self-limiting dengue fever that can progress to a life-threatening syndrome called dengue hemorrhagic fever (DHF) [23]. DENV life cycle resembles to that of HCV, with a few modifications, that include involvement of a mosquito vector and infection of different host cells (monocytes and macrophages). Upon getting bitten with an infected mosquito, DENV initially infects skin cells and subsequently immune cells of the neighboring lymph node, such as macrophages and monocytes. DENV is internalized by receptor-mediated endocytosis through association with the mannose receptor, C-type lectin DC-SIGN, and the CLEC5A, and upon fusion with endosomes, the single copy of viral RNA is released into the cytoplasm [24, 25]. The viral genome of ~11,000 bases encodes a single large polyprotein that gets cleaved by cellular and viral proteases at specific recognition sites into three structural and seven nonstructural proteins. The structural proteins are capsid (C) protein, membrane (prM) protein, and envelope (E) glycoprotein. The nonstructural proteins include NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5 [26]. The E glycoprotein mediates attachment of the virion to the receptor and facilitates fusion between viral and host cell membrane. The NS5 is the RNA polymerase responsible for viral RNA replication. Replication of DENV occurs in close apposition with virus-induced extensively rearranged ER membrane,

called replication complex (RC). RCs contain viral RNA, viral proteins, and host cell factors. The newly synthesized viral RNA is encapsidated by the C protein that emerges toward the ER lumen, thereby acquiring a lipid bilayer envelope containing prM and E proteins. These immature virions traffic along the secretory pathway to the Golgi, where prM is cleaved into mature pr peptide and virion-associated M by host cell furin proteases, thereby generating mature infectious virions that are secreted [27].

Similar to the HCV core, the DENV C protein is released from the polyprotein by two subsequent peptidase cleavages: first by signal peptidase and second by the viral NS3/NS2B protease [28]. The C protein localizes to the periphery of LDs, and this association is critical for viral replication, since mutation in the hydrophobic residues that reside within the $\alpha 2$ helix of the C protein impairs C protein localization and disrupts viral replication [29]. Intriguingly, the viral NS3 protein interacts with host cell fatty acid synthase enzyme and recruits it to the viral replication machinery. As a result, inhibiting fatty acid synthase attenuates production of LDs and thereby impairs assembly and release of infectious virions [29]. Thus, LDs play an essential role in completing steps of DENV replication and maturation.

Rotavirus—Rotaviruses (RVs) are non-enveloped double-stranded RNA viruses that belong to *Reoviridae* family. RVs infect enterocytes and are a major cause of acute gastroenteritis in infants and young children globally. Despite the fact that an effective RV vaccine is in use in >100 countries that have reduced associated disease, most of the RV infection-related mortality occurs mainly in countries with low socioeconomic conditions [30]. The RV genome encodes six structural proteins (VP1–VP4, VP6, and VP7) and six nonstructural proteins (NSP1–NSP6). Usually the RVs exist as triple-layered particles (TLPs) comprising four major capsid proteins (VP2, VP4, VP6, and VP7). Intestinal cells internalize triple-layered RVs via receptor-mediated endocytosis, where outer layer of RVs is removed inside the endocytic vesicles. Subsequently double-layered particles (DLPs) actively transcribe the viral RNA into mRNA that is released into the cytoplasm. Early steps of viral assembly occur inside intracellular inclusion bodies, called viroplasms (vpls) or “viral factories” that are induced by RVs. The early viral morphogenesis, and replication, of viral RNA to produce dsRNA occurs inside vpls. Partially assembled DLPs are released from the vpls into the ER lumen, where they acquire outer layer of the rough ER, and mature as infectious TLP-RVs that traffics to the Golgi and are released via the exocytotic pathway.

Association between vpls and LDs has been reported where LDs might serve as a platform for viral packaging, as demonstrated by colocalization of lipophilic dyes, and LD marker proteins, Plin1 and Plin2, with NSP2 and NSP5, two key viral proteins that are crucial for vpl formation [31, 32]. Viroplasm-mediated factors, such as NSP5, appear to regulate the recruitment of LD components to vpl during early RVs replication cycle. siRNA directed against NSP5 prevents the localization of Plin1 with NSP5 [31]. In addition, viral NSP4 colocalizes with caveolin-1 [33] that localizes to droplet surface and has important function in regulating LD biogenesis and degradation [34]. Interestingly, dsRNA of RVs, together with vpl protein NSP5, and LD-resident protein Plin2 were found to co-enrich in low-density gradient

fractions isolated from rotavirus-infected cell extracts, further reinforcing the notion that LDs form complexes with vpls [31]. Upon comparing the lipidomes of RV-infected and uninfected MA104 kidney cells, it was found that concentration of almost all lipids was elevated in RV-infected cells. Employing mass spectrometry, the lipids were segregated into 14 different classes. Low-density gradient fractions, containing peaks of the RVs dsRNA genome, LDs, and vpls-associated proteins, were also found to be enriched with lipids that are typically found in LDs, suggesting close relationship of LDs with vpls [35]. Inhibiting LD biogenesis by using triacsin C impairs RV replication, vpls formation, and viral-induced cell death in cultured cells, further implying that RV proteins establish close relationship with LDs during RV life cycle for its propagation [31]. Similarly, upon induction of lipolysis by using isoproterenol (beta-adrenergic stimulator) and isobutylmethylxanthine (IBMX, phosphodiesterase inhibitor), both of which upregulate intracellular cAMP levels, thereby activating hormone-dependent lipases, leading to dispersion and regression of LDs which resulted in inhibition of RV replication and reduced rate of infection [31]. These findings suggest that LDs play a vital role in the pathogenesis of RVs, and hence better characterizing role of LDs in the replication of RVs will provide new therapeutic intervention opportunities.

23.3 LDs and Bacterial Infection

Many intracellular bacteria target host LDs for their survival and facilitating infection (reviewed in [36]). In this section we discuss the role of LDs during *Chlamydia trachomatis* and *Mycobacterium* spp. infection.

C. trachomatis is a Gram-negative, obligate intracellular bacterium that is sexually transmitted and causes chronic disease of the urogenital tract, while other *Chlamydia* spp. infect the eye epithelium and cause trachoma, a leading cause of non-congenital blindness globally. Though these infections can be treated with antibiotics, there is lack of efficacious vaccines; thus it is imperative to better understand the pathogenesis of *C. trachomatis*. The bacterial life cycle comprises two phases: the first being an extracellular, infectious, environmentally stable, metabolically inactive, and inert phase called elementary bodies (EBs) and the second being replicative, metabolically active phase called reticulate bodies (RBs) [37]. The metabolically inactive EBs infect host mucosal cells, escape phagolysosomal fusion, and establish a membrane-bound parasitophorous vacuole (PVs), also known as inclusions, that are mostly inaccessible to the trafficking machinery of the host cell, thereby generating a niche for the intracellular survival of *Chlamydia*. Inside PVs, EBs differentiate into RBs, and upon several rounds of replication, RBs once again get converted to EBs that are released from the host cells to infect neighboring cells.

Previous studies have reported importance of host LDs for growth and propagation of *C. trachomatis*. *C. trachomatis* infection upregulates host LD accumulation and increased cholesterol ester (CE) production in cultured mammalian cells [38]. During early stages of infection, host LDs are reorganized so as to surround

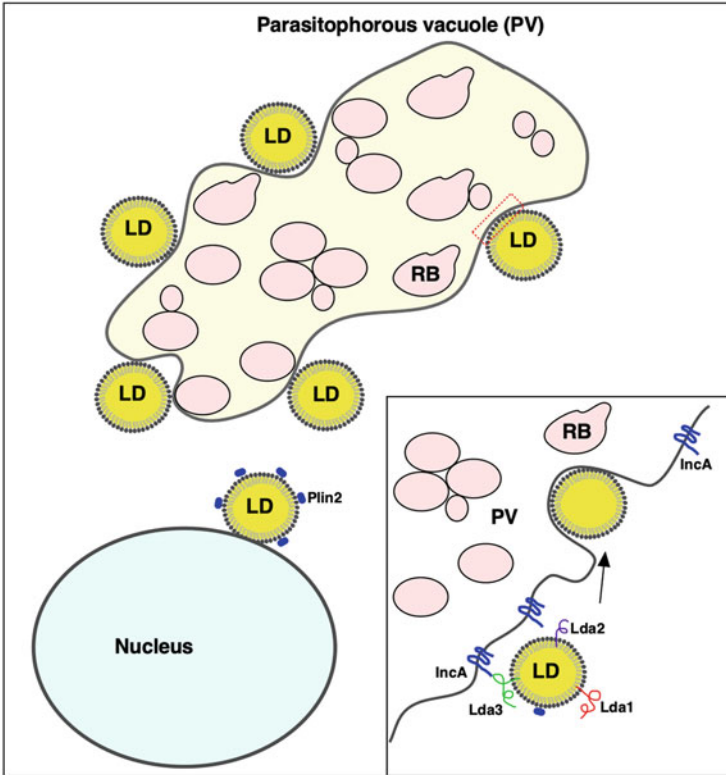


Fig. 23.2 *Chlamydia* inclusions interact with host LDs. Replication of *Chlamydia* occurs inside the membrane-bound parasitophorous vacuole (PV). Host LDs are recruited to the periphery of PV membrane. Secreted bacterial Lda proteins have affinity to the monolayer surface of LDs, thereby resulting in displacement of native LD-associated coat proteins. Interaction of Lda3 on LD surface with IncA, a PV membrane protein, results in the internalization of host LDs inside the inclusion. Boxed region is shown in the inset

the periphery of the PV membrane and eventually LDs being internalized inside the PV to provide nutrients and lipids to the replicating bacteria (Fig. 23.2).

During *C. trachomatis* multiplication, it employs several proteins to hijack host LDs to salvage lipid substrates for its propagation. The bacterial genome encodes three key proteins, known as LD-associated proteins, Lda1, Lda2, and Lda3, that are translocated into the host cytoplasm to target LD periphery [38]. Host LDs decorated with Lda proteins results in mislocalization of native LD surface proteins, such as Plin2, an LD coat protein, thereby assisting in lipolysis of stored LD contents (Fig. 23.2). IncA, an inclusion membrane protein of *Chlamydia*, delineates the PV membrane and was found to transiently associate with LDs [39]. Interaction of Lda3-coated host LDs with IncA at the PV membrane results in translocation of LDs inside the inclusion, suggesting IncA might play a crucial role in the translocation of host

LDs [39] (Fig. 23.2). CT149, a cholesterol esterase expressed by *C. trachomatis*, may act upon LDs inside the PV to hydrolyze stored TGs and CE, thereby freeing free fatty acids and cholesterol to be utilized by the bacteria [40]. Inhibition of LD formation by treatment with triacsin C resulted in smaller inclusions and decreased bacterial growth, suggesting LDs support intracellular growth and infection of the bacteria [38]. Findings from a number of studies support the notion that a strong interaction between *C. trachomatis* and host LDs is critical for bacterial pathogenesis; however, the protein composition of the junctional interphase between LDs and PV membrane is not completely known. What host factors are crucial for establishing a successful hijack of host LDs for the benefit of the bacteria remains to be determined.

Mycobacterium spp. is another obligate intracellular bacterial pathogen that causes tuberculosis and leprosy diseases globally. LD-filled macrophages are the hallmark feature of *M. tuberculosis* (*Mtb*) and *M. leprae* infections. *Mtb* infects primary alveolar macrophages in the respiratory tract and causes severe pulmonary disease tuberculosis (TB). *Mtb* is transmitted by aerosol droplets. Inside macrophages, *Mtb* evades host phagolysosomal fusion and matures inside the phagosome so as to escape host immune response-mediated clearance [41]. In this state, *Mtb* can persist latently for many years without showing clinical signs of TB, until activated by weakened immune system of the host. This dormancy-like phenotype results in granuloma formation, a structure that includes infected macrophages surrounded by “foamy” LD-laden macrophages, mononuclear phagocytes, and T lymphocytes inside a fibrous layer of endothelial cells [42, 43]. *Mtb* manipulates host lipid metabolism through modulation of PPAR nuclear receptors, thereby resulting in induction of LD-filled foamy macrophages for its persistence within the host [36, 44]. *Mtb* utilizes TG as its main source of energy. *Mtb* migrate toward host LDs in foamy macrophages, eventually leading to engulfment of these LDs and accumulating the lipids within. *Mtb* genome encodes a lipase, LipY, a homolog of mammalian HSL, that can degrade host TG to liberate fatty acids for the utilization by the bacterium [45]. *Mtb* can also synthesize its own TG by the enzyme Tgs1 (triacylglycerol synthase 1), where host-derived fatty acids act as substrate for Tgs1 [46]. In addition, *Mtb* encodes for a cholesterol importer protein machinery, mce4, enabling *Mtb* to derive host cholesterol for nutritional purpose, and it appears to be essential for long-term bacterial persistence [47]. Thus, current studies suggest that host LDs serve as an energy reservoir for lasting survival of *Mtb*. Further investigations are needed to better characterize intricate host-bacterial interactions to further elucidate mechanisms of bacterial pathogenesis.

M. leprae causes leprosy, a granulomatous disease of the upper respiratory tract, that primarily affects the skin and the peripheral nerves. *M. leprae* invades the Schwann and glial cells of the peripheral nerves that result in demyelination and nerve fibrosis which if left untreated can lead to irreversible nerve dysfunction that provokes deformities and physical disabilities. Infection of *M. leprae* manifests into two major diseases: the one includes tuberculoid leprosy resulting in the formation of granuloma and Schwann cell death, whereas the second form called lepromatous leprosy is characterized by the presence of LD-filled foam cell-containing lesions

[36, 48]. The main symptoms of leprosy are formation of deformities, pale-colored skin sores, and lumps that persist for a long duration. During infection, LD formation was upregulated in murine and human macrophages, and expression of LD coat proteins, Plin1 and Plin2, were found to be elevated that colocalized with bacterium-containing phagosome. A similar redistribution of host LDs occurs in Schwann cells where Plin2-coated LDs colocalize with bacterial phagosome in cultured cells and nerve biopsies taken from patients [49–51]. Redistribution of LDs is dependent on PI3K signaling and is mediated by rearrangement of cytoskeletal factors [50]. Interference with motility of LDs by blocking cytoskeletal functioning resulted in impaired bacterial survival in infected cells. Induction of LD formation, upregulation of Plin proteins, and subsequently production of prostaglandins, an anti-inflammatory mediator, are dependent of TLR2 (toll-like receptor 2) signaling. Toll-like receptors play a crucial role in the recognition of pathogens and activation of innate immunity. The mechanistic details of how LDs assist in *M. leprae* growth are not completely understood. Accumulation of LDs gives foamy appearance to lepromatous leprosy lesions, where LDs might provide nutrients to the pathogen and substrate for innate immunity response (reviewed in [52]). Surprisingly, *M. leprae* infection downregulated expression of HSL, thereby inhibiting lipolysis and facilitating LD accumulation. In agreement with this, macrophages treated with clofazimine, an antibiotic given to leprosy patients, resulted in reduced Plin2 levels and increased HSL level in cultured cells and clinical samples, therefore leading to reduced LD accumulation and inhibiting *M. leprae* growth, suggesting that LD accumulation by inhibiting lipolysis is an important step for persistent pathogen multiplication [53]. Taken together, available studies support a model in which mycobacterial infections result in establishment of close interactions with host LDs at the interphase of evading host immune response and obtaining nutrients for its long-lasting survival.

23.3.1 LDs and Parasite Infection

LDs accumulate in the cytoplasm of various cell types as a response to infection with protozoan parasites such as *Trypanosoma cruzi*, *Toxoplasma gondii*, and *Plasmodium* spp. Interaction of parasites with the host modulates the lipid metabolism of both organisms. Some parasites reside inside their parasitophorous vacuole (PV), while other escape and multiply in host cytoplasm. PV allows the parasite to grow without fusing with the phagolysosomes of the host cell. Intracellular parasites have evolved ways to exploit host LDs for their own benefit and long-term survival. Host LDs serve as a source of nutrients, such as lipids for membrane synthesis of the parasite, a crucial step for parasite growth, and to evade host immune response (reviewed in [54]). Protozoan parasites cannot produce cholesterol de novo, despite having complex machinery to interact with their host. Thus, they derive cholesterol from LDL or host LDs. *T. cruzi*, the etiologic agent of Chagas disease, invades the host cells via LDL receptor (LDLr), resides inside PV, and facilitates subsequent fusion of the PV with the host lysosomal compartment to egress from the PV into the

cytoplasm of the host cell where it further replicates [55, 56]. *T. cruzi* infection induces LDLr levels; therefore disruption of LDLr reduces the total parasite burden in infected cells [55]. On the other hand, *T. gondii* and *Plasmodium* spp. build a PV that is permissive for the growth of the pathogen, and it does not fuse with the host lysosome.

T. gondii infection leads to toxoplasmosis that can cause serious complications in pregnant females and people with compromised immune system. Infection of *T. gondii* in host cells results in an induction of LD formation together with an upregulation of LD-associated proteins, Plin2, and DGAT2. Throughout the infection, host LDs were found to surround the PV membrane. *T. gondii* inside the PV interacts with host LDs to fulfill its nutritional needs by synthesizing a projection of nanotubular structures called intravacuolar network (IVN) of membranes [57–59]. *T. gondii* intercepts and engulfs Rab7-coated host LDs into the PV lumen by unknown mechanisms that are degraded by the parasite lipases to liberate neutral lipids from host LDs to support the pathogen growth [58]. Host cells lacking LDs, or having impaired TG lipolysis, show reduced *T. gondii* multiplication and infection. Similarly, impairment in IVN formation displayed reduced lipid uptake capacity from host LDs, highlighting the crucial role of LDs in pathogen replication cycle [58].

Plasmodium falciparum infection causes the deadliest form of malaria. It is transmitted by the bite of an infected *Anopheles* mosquito. Sporozoites released in blood infect hepatocytes, where they undergo asexual replication to produce merozoites. Merozoites infect mature erythrocytes (RBCs), replicate and multiply, and lead to rupture of the infected RBCs that causes high fever and chills in patients. Mature RBCs lack any internal organelles or the machinery for protein and lipid synthesis. Hemoglobin, the main constituent of the mature RBCs, is degraded by the parasite and used as an amino acid source. Increased TG content and LDs have been detected in infected RBCs [60]. Remarkably, the parasite encodes its own DGAT enzyme to synthesize TG-rich LDs that are expressed specifically in the intra-erythrocytic stage and are important for growth and proliferation of the parasite [60]. However, the parasite does not express any lipase to degrade TG, and the accumulating LDs in the PV are believed to be involved in the detoxification of heme, further implying important role LDs play in the replication of the pathogen inside host cells [61].

23.3.2 Novel Approaches to Probe Host-Pathogen Interactions

Compared to traditional methods of lipid analysis such as enzyme immunoassays (EIAs) and thin-layer chromatography (TLC), more advanced modern methods including gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS), nuclear magnetic resonance (NMR) spectroscopy, column chromatography, and microfluidic devices are now being employed in the field of lipidomics to identify, characterize, quantify, and elucidate the structure and function of diverse lipids species [62]. Lipidomics is a newly

emerged discipline of metabolomics for the characterization and quantification of molecular lipid species in total lipid extracts. Charged ionized analytes can be separated by their mass-to-charge (m/z) ratio, and structural information for components in complex mixtures can be obtained by performing fragmentation of lipid ions by collision-induced dissociation (CID). This technique is called tandem MS, or MS/MS. “Shotgun lipidomics” allows identification and quantification of thousands of cellular lipid species in a high-throughput untargeted manner. On the other hand, targeted lipidomics using LC-MS and LC-MS/MS approach allows to detect and quantify a particular few classes of lipids of interest. LC-MS-based lipidomic analyses of monocyte/macrophage covering membrane lipids (sterols, glycerophospholipids, and sphingolipids), and signaling lipids (eicosanoids, phosphoinositides), have revealed sophisticated metabolic networks during the differentiation process of macrophages, TLR activation, and interaction between host and pathogens [63–65]. Similarly, lipidome of the pathogen such as *Mtb* has been established [66] and has provided new leads to study lipid metabolism during the complex host-pathogen interaction. A limitation of MS-based lipidomic approach is that it lacks the spatial resolution to identify and determine localized enrichment of particular lipid species during the pathogen infection; however, newly discovered MS-based imaging approach now provides spatial resolution of metabolites during the pathogen infection [67, 68]. Combining lipidomics with the usage of immunofluorescent lipids and lipid-binding probes for microscopy will enhance the spatial resolution of host-pathogen interaction and will aid in dissecting the role of lipids in the pathogenesis and disease manifestation. Future research employing a systems-level approach by combining proteomics and genomics together with lipidomics will broaden our understating of the complex interplay between host and pathogen and provide new insights into the interconnected lipid metabolic network for the identification of potential lipid or metabolic pathway for therapeutic interventions.

23.4 Concluding Remarks

During the last few decades, our current understanding about LD biology has greatly expanded, with novel insights into how droplets are assembled, turned over, and are interconnected with other cellular organelles at membrane contact sites. Several discoveries have uncovered key proteins that play crucial role in establishing regulated LD formation. The intriguing finding that pathogens, such as viruses, bacteria, and parasites, hijack LDs for their benefit highlight an important and previously unrecognized function of LDs. Though it is well established that LDs play an important role in pathogen multiplication and disease manifestation, we have just begun to scratch the surface of LD-pathogen interaction as numerous sequences of events are involved during which LDs and lipids in general play instrumental role. Several key questions need further investigations: Do assembly of infectious pathogen occur at specialized ER subdomains at which nascent LDs are born? Are there significant differences between how LDs interact with different pathogens? What is

the molecular mechanism of induction of LD formation, is it pathogen-driven or host-driven? During pathogen infection how do repertoire of LD surface proteins change? Which host/pathogen factors are responsible for establishing close contact with the LDs? Ultimately a clearer elucidation of pathways implicated in disease progression will pave the way forward for better understanding of the role of lipid metabolism in disease progression aimed at identification of biomarkers and developing therapeutic approaches.

Declaration of Competing Interests

The authors declare no competing interests.

Acknowledgments This work was supported by the Early Career Intramural Project of the All India Institute of Medical Sciences (AIIMS), New Delhi. A. Bhattacharyya is supported by Ramanujan Fellowship of Science & Engineering Research Board (SERB), Department of Science and Technology (DST), Govt. of India (grant: RJF/2019/000040).

References

1. Olzmann JA, Carvalho P (2019) Dynamics and functions of lipid droplets. *Nat Rev Mol Cell Biol* 20:137–155
2. Choudhary V, Schneider R (2020) Lipid droplet biogenesis from specialized ER subdomains. *Microb Cell* 7:218–221
3. Ben M'barek K, Ajjaji D, Chorlay A, Vanni S, Foret L, Thiam AR (2017) ER membrane phospholipids and surface tension control cellular lipid droplet formation. *Dev Cell* 41:591–604.e7
4. Choudhary V, EL Atab O, Mizzon G, Prinz WA, Schneider R (2020) Seipin and Nem1 establish discrete ER subdomains to initiate yeast lipid droplet biogenesis. *J Cell Biol* 219:e201910177
5. Choudhary V, Golani G, Joshi AS, Cottier S, Schneider R, Prinz WA, Kozlov MM (2018) Architecture of lipid droplets in endoplasmic reticulum is determined by phospholipid intrinsic curvature. *Curr Biol* 28:915–926.e9
6. Santinho A, Salo VT, Chorlay A, Li S, Zhou X, Omrane M, Ikonen E, Thiam AR (2020) Membrane curvature catalyzes lipid droplet assembly. *Curr Biol* 30:2481–2494.e6
7. Chung J, Wu X, Lambert TJ, Lai ZW, Walther TC, Farese RV (2019) LDAH1 and Seipin form a lipid droplet assembly complex. *Dev Cell* 51:551–563.e7
8. Haemmerle G, Lass A, Zimmermann R, Gorkiewicz G, Meyer C, Rozman J, Heldmaier G, Maier R, Theussl C, Eder S, Kratky D, Wagner EF, Klingenspor M, Hoefler G, Zechner R (2006) Defective lipolysis and altered energy metabolism in mice lacking adipose triglyceride lipase. *Science* 312:734–737
9. Vaughan M, Berger JE, Steinberg D (1964) Hormone-sensitive lipase and Monoglyceride lipase activities in adipose tissue. *J Biol Chem* 239:401–409
10. Guijas C, Rodriguez JP, Rubio JM, Balboa MA, Balsinde J (2014) Phospholipase A2 regulation of lipid droplet formation. *Biochim Biophys Acta* 1841:1661–1671
11. Cermelli S, Guo Y, Gross SP, Welte MA (2006) The lipid-droplet proteome reveals that droplets are a protein-storage depot. *Curr Biol* 16:1783–1795
12. Fei W, Wang H, Fu X, Bielby C, Yang H (2009) Conditions of endoplasmic reticulum stress stimulate lipid droplet formation in *Saccharomyces cerevisiae*. *Biochem J* 424(1):61–67
13. Herker E, Ott M (2012) Emerging role of lipid droplets in host/pathogen interactions. *J Biol Chem* 287:2280–2287
14. Ploss A, Evans MJ (2012) Hepatitis C virus host cell entry. *Curr Opin Virol* 2:14–19

15. Paul D, Madan V, Bartenschlager R (2014) Hepatitis C virus RNA replication and assembly: living on the fat of the land. *Cell Host Microbe* 16:569–579
16. Vieyres G, Pietschmann T (2019) HCV pit stop at the lipid droplet: refuel lipids and put on a lipoprotein coat before exit. *Cell* 8
17. Herker E, Harris C, Hernandez C, Carpentier A, Kaehlcke K, Rosenberg AR, Farese RV, Ott M (2010) Efficient hepatitis C virus particle formation requires diacylglycerol acyltransferase-1. *Nat Med* 16:1295–1298
18. Miyanari Y, Atsuzawa K, Usuda N, Watashi K, Hishiki T, Zayas M, Bartenschlager R, Wakita T, Hijikata M, Shimotohno K (2007) The lipid droplet is an important organelle for hepatitis C virus production. *Nat Cell Biol* 9:1089–1097
19. Kim D, Goo JI, Kim MI, Lee SJ, Choi M, Than TT, Nguyen PH, Windisch MP, Lee K, Choi Y, Lee C (2018) Suppression of hepatitis C virus genome replication and particle production by a novel Diacylglycerol Acyltransferases inhibitor. *Molecules* 23
20. Boulant S, Douglas MW, Moody L, Budkowska A, Targett-Adams P, Mclauchlan J (2008) Hepatitis C virus core protein induces lipid droplet redistribution in a microtubule- and dynein-dependent manner. *Traffic* 9:1268–1282
21. Huang H, Sun F, Owen DM, Li W, Chen Y, Gale M, Ye J (2007) Hepatitis C virus production by human hepatocytes dependent on assembly and secretion of very low-density lipoproteins. *Proc Natl Acad Sci U S A* 104:5848–5853
22. Wrensch F, Crouchet E, Ligat G, Zeisel MB, Keck ZY, Fong SKH, Schuster C, Baumert TF (2018) Hepatitis C virus (HCV)-Apolipoprotein interactions and immune evasion and their impact on HCV vaccine design. *Front Immunol* 9:1436
23. Rodenhuis-Zybert IA, Wilschut J, Smit JM (2010) Dengue virus life cycle: viral and host factors modulating infectivity. *Cell Mol Life Sci* 67:2773–2786
24. Miller JL, DE Wet BJ, Martinez-Pomares L, Radcliffe CM, Dwek RA, Rudd PM, Gordon S (2008) The mannose receptor mediates dengue virus infection of macrophages. *PLoS Pathog* 4:e17
25. Van Der Schaar HM, Rust MJ, Chen C, van der Ende-Metselaar H, Wilschut J, Zhuang X, Smit JM (2008) Dissecting the cell entry pathway of dengue virus by single-particle tracking in living cells. *PLoS Pathog* 4:e1000244
26. Clyde K, Kyle JL, Harris E (2006) Recent advances in deciphering viral and host determinants of dengue virus replication and pathogenesis. *J Virol* 80:11418–11431
27. Zybert IA, Van Der Ende-Metselaar H, Wilschut J, Smit JM (2008) Functional importance of dengue virus maturation: infectious properties of immature virions. *J Gen Virol* 89:3047–3051
28. Niyomrattanakit P, Winoyanu wattikun P, Chanprapaph S, Angsuthanasombat C, Panyim S, Katzenmeier G (2004) Identification of residues in the dengue virus type 2 NS2B cofactor that are critical for NS3 protease activation. *J Virol* 78:13708–13716
29. Samsa MM, Mondotte JA, Iglesias NG, Assuncao-Miranda I, Barbosa-Lima G, DA Poian AT, Bozza PT, Gamarnik AV (2009) Dengue virus capsid protein usurps lipid droplets for viral particle formation. *PLoS Pathog* 5:e1000632
30. Desselberger U (2014) Rotaviruses. *Virus Res* 190:75–96
31. Cheung W, Gill M, Esposito A, Kaminski CF, Courousse N, Chwetzoff S, Trugnan G, Keshavan N, Lever A, Desselberger U (2010) Rotaviruses associate with cellular lipid droplet components to replicate in viroplasms, and compounds disrupting or blocking lipid droplets inhibit viroplasm formation and viral replication. *J Virol* 84:6782–6798
32. Fabbretti E, Afrikanova I, Vascotto F, Burrone OR (1999) Two non-structural rotavirus proteins, NSP2 and NSP5, form viroplasm-like structures in vivo. *J Gen Virol* 80(Pt 2):333–339
33. Parr RD, Storey SM, Mitchell DM, Mcintosh AL, Zhou M, Mir KD, Ball JM (2006) The rotavirus enterotoxin NSP4 directly interacts with the caveolar structural protein caveolin-1. *J Virol* 80:2842–2854
34. Cohen AW, Razani B, Schubert W, Williams TM, Wang XB, Iyengar P, Brasaemle DL, Scherer PE, Lisanti MP (2004) Role of caveolin-1 in the modulation of lipolysis and lipid droplet formation. *Diabetes* 53:1261–1270

35. Gaunt ER, Zhang Q, Cheung W, Wakelam MJO, Lever AML, Desselberger U (2013) Lipidome analysis of rotavirus-infected cells confirms the close interaction of lipid droplets with viroplasms. *J Gen Virol* 94:1576–1586
36. Libbing CL, Mcdevitt AR, Azcueta RP, Ahila A, Mulye M (2019) Lipid droplets: a significant but understudied contributor of host(–)bacterial interactions. *Cell* 8:354
37. Belland R, Ojcius DM, Byrne GI (2004) Chlamydia. *Nat Rev Microbiol* 2:530–531
38. Kumar Y, Cocchiari J, Valdivia RH (2006) The obligate intracellular pathogen chlamydia trachomatis targets host lipid droplets. *Curr Biol* 16:1646–1651
39. Cocchiari JL, Kumar Y, Fischer ER, Hackstadt T, Valdivia RH (2008) Cytoplasmic lipid droplets are translocated into the lumen of the chlamydia trachomatis parasitophorous vacuole. *Proc Natl Acad Sci U S A* 105:9379–9384
40. Peters J, Onguri V, Nishimoto SK, Marion TN, Byrne GI (2012) The chlamydia trachomatis CT149 protein exhibits esterase activity in vitro and catalyzes cholesteryl ester hydrolysis when expressed in HeLa cells. *Microbes Infect* 14:1196–1204
41. Weber SS, Ragaz C, Hilbi H (2009) Pathogen trafficking pathways and host phosphoinositide metabolism. *Mol Microbiol* 71:1341–1352
42. Sandoz KM, Valiant WG, Eriksen SG, Hruba DE, Allen RD, Rockey DD (2014) The broad-spectrum antiviral compound ST-669 restricts chlamydial inclusion development and bacterial growth and localizes to host cell lipid droplets within treated cells. *Antimicrob Agents Chemother* 58:3860–3866
43. Zahrt TC (2003) Molecular mechanisms regulating persistent Mycobacterium tuberculosis infection. *Microbes Infect* 5:159–167
44. Almeida PE, Carneiro AB, Silva AR, Bozza PT (2012) PPAR γ expression and function in mycobacterial infection: roles in lipid metabolism, immunity, and bacterial killing. *PPAR Res* 2012:383829
45. Deb C, Daniel J, Sirakova TD, Abomoelak B, DUBEY VS, Kolattukudy PE (2006) A novel lipase belonging to the hormone-sensitive lipase family induced under starvation to utilize stored triacylglycerol in Mycobacterium tuberculosis. *J Biol Chem* 281:3866–3875
46. Daniel J, Maamar H, Deb C, Sirakova TD, Kolattukudy PE (2011) Mycobacterium tuberculosis uses host triacylglycerol to accumulate lipid droplets and acquires a dormancy-like phenotype in lipid-loaded macrophages. *PLoS Pathog* 7:e1002093
47. Pandey AK, Sassetti CM (2008) Mycobacterial persistence requires the utilization of host cholesterol. *Proc Natl Acad Sci U S A* 105:4376–4380
48. Scollard DM, Joyce MP, Gillis TP (2006) Development of leprosy and type 1 leprosy reactions after treatment with infliximab: a report of 2 cases. *Clin Infect Dis* 43:e19–e22
49. Mattos KA, D'ávila H, Rodrigues LS, Oliveira VG, Sarno EN, Atella GC, Pereira GM, Bozza PT, Pessolani MC (2010) Lipid droplet formation in leprosy: toll-like receptor-regulated organelles involved in eicosanoid formation and mycobacterium leprae pathogenesis. *J Leukoc Biol* 87:371–384
50. Mattos KA, Lara FA, Oliveira VG, Rodrigues LS, D'ávila H, Melo RC, Manso PP, Sarno EN, Bozza PT, Pessolani MC (2011) Modulation of lipid droplets by mycobacterium leprae in Schwann cells: a putative mechanism for host lipid acquisition and bacterial survival in phagosomes. *Cell Microbiol* 13:259–273
51. Tanigawa K, Suzuki K, Nakamura K, Akama T, Kawashima A, Wu H, Hayashi M, Takahashi S, Ikuyama S, Ito T, Ishii N (2008) Expression of adipose differentiation-related protein (ADRP) and perilipin in macrophages infected with mycobacterium leprae. *FEMS Microbiol Lett* 289:72–79
52. De Mattos KA, Sarno EN, Pessolani MC, Bozza PT (2012) Deciphering the contribution of lipid droplets in leprosy: multifunctional organelles with roles in mycobacterium leprae pathogenesis. *Mem Inst Oswaldo Cruz* 107(Suppl 1):156–166
53. Degang Y, Akama T, Hara T, Tanigawa K, Ishido Y, Gidoh M, Makino M, Ishii N, Suzuki K (2012) Clofazimine modulates the expression of lipid metabolism proteins in mycobacterium leprae-infected macrophages. *PLoS Negl Trop Dis* 6:e1936

54. Vallochi AL, Teixeira L, Oliveira KDS, Maya-Monteiro CM, Bozza PT (2018) Lipid droplet, a key player in host-parasite interactions. *Front Immunol* 9:1022
55. Nagajyothi F, Weiss LM, Silver DL, Desruisseaux MS, Scherer PE, Herz J, Tanowitz HB (2011) *Trypanosoma cruzi* utilizes the host low density lipoprotein receptor in invasion. *PLoS Negl Trop Dis* 5:e953
56. Tardieux I, Webster P, Ravesloot J, Boron W, Lunn JA, Heuser JE, Andrews NW (1992) Lysosome recruitment and fusion are early events required for trypanosome invasion of mammalian cells. *Cell* 71:1117–1130
57. Mercier C, Dubremetz JF, Rauscher B, Lecordier L, Sibley LD, Cesbron-Delauw MF (2002) Biogenesis of nanotubular network in toxoplasma parasitophorous vacuole induced by parasite proteins. *Mol Biol Cell* 13:2397–2409
58. Nolan SJ, Romano JD, Coppens I (2017) Host lipid droplets: an important source of lipids salvaged by the intracellular parasite *Toxoplasma gondii*. *PLoS Pathog* 13:e1006362
59. Romano JD, Coppens I (2013) Host organelle hijackers: a similar modus operandi for *Toxoplasma gondii* and *Chlamydia trachomatis*: co-infection model as a tool to investigate pathogenesis. *Pathog Dis* 69:72–86
60. Vielemeyer O, Mcintosh MT, Joiner KA, Coppens I (2004) Neutral lipid synthesis and storage in the intraerythrocytic stages of *Plasmodium falciparum*. *Mol Biochem Parasitol* 135:197–209
61. Jackson KE, Klonis N, Ferguson DJ, Adisa A, Dogovski C, Tilley L (2004) Food vacuole-associated lipid bodies and heterogeneous lipid environments in the malaria parasite, *Plasmodium falciparum*. *Mol Microbiol* 54:109–122
62. Teng O, Ang CKE, Guan XL (2017) Macrophage-bacteria interactions—a lipid-centric relationship. *Front Immunol* 8:1836
63. Koberlin MS, Snijder B, Heinz LX, Baumann CL, Fauster A, Vladimer GI, Gavin AC, Superti-Furga G (2015) A conserved circular network of Coregulated lipids modulates innate immune responses. *Cell* 162:170–183
64. Tam VC (2013) Lipidomic profiling of bioactive lipids by mass spectrometry during microbial infections. *Semin Immunol* 25:240–248
65. Wenk MR (2006) Lipidomics of host-pathogen interactions. *FEBS Lett* 580:5541–5551
66. Crick PJ, Guan XL (2016) Lipid metabolism in mycobacteria—insights using mass spectrometry-based lipidomics. *Biochim Biophys Acta* 1861:60–67
67. Barcelo-Coblijn G, Fernandez JA (2015) Mass spectrometry coupled to imaging techniques: the better the view the greater the challenge. *Front Physiol* 6:3
68. Hulme HE, Meikle LM, Wessel H, Strittmatter N, Swales J, Thomson C, Nilsson A, Nibbs RJB, Milling S, Andren PE, Mackay CL, Dexter A, Bunch J, Goodwin RJA, Burchmore R, Wall DM (2017) Mass spectrometry imaging identifies palmitoylcarnitine as an immunological mediator during salmonella *Typhimurium* infection. *Sci Rep* 7:2786



Lipid Structure, Function, and Lipidomic Applications

24

Khusboo Arya, Sana Akhtar Usmani, Nitin Bhardwaj,
Sudhir Mehrotra, and Ashutosh Singh

Abstract

Lipids are a key component of membranes and act as bioactive signaling molecules in biological systems. Over the years, lipids have gained much significance as their roles are being highlighted through various studies. The structural and functional diversity of lipid structure allows great variety to roles that these lipids can play in cellular organisms. With an increase of research interests in area related to lipids, much emphasis has been put toward the manner in which these can be analyzed. Recent developments in mass spectrometry-based platforms have allowed us to generate global lipid maps at the cellular level and elucidate the molecular complexity of lipid structures in detail. In this chapter we discuss some aspects of lipid classification, their structure and functions, and mass spectrometry-based lipidomics approach used for determination of lipid compositions.

Keywords

Lipids · Lipidomics · Mass spectrometry

Khusboo Arya and Sana Akhtar Usmani contributed equally with all other contributors.

K. Arya · S. A. Usmani · S. Mehrotra · A. Singh (✉)

Department of Biochemistry, University of Lucknow, Lucknow, Uttar Pradesh, India

e-mail: singh_ashutosh@lkouniv.ac.in

N. Bhardwaj

Department of Zoology and Environmental Science, Gurukula Kangri Vishwavidyalaya, Haridwar, Uttarakhand, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

441

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*,
https://doi.org/10.1007/978-981-16-0691-5_24

List of Abbreviations

GlcCer	glucosylceramide
PUFAs	polyunsaturated fatty acids
SGs	steryl glucosides
FA	fatty acids
AM	arbuscular mycorrhizal
Q-TOF	quadrupole time of flight
nESI	nano-electrospray
Q1	quadrupole 1
q2	quadrupole 2
Q3	quadrupole 3

24.1 Introduction to Lipids

Lipids are nonpolar molecules and soluble usually in nonpolar solvents. These molecules generally comprise oils and fats and yield high energy for various cellular functions [1]. Generally, lipids are composed for a nonpolar hydrocarbon chain along with oxygen-containing polar region [1]. In humans, these are stored in the adipose tissue [2]. Lipids can be classified in many ways like on their polarity complexity, attached molecules or their function, and hydrolyzing property. When they cannot be hydrolyzed into smaller molecules, they are considered as non-saponifiable lipids like cholesterol, prostaglandins, etc. Lipids which are hydrolyzed due to the presence of one or more ester groups are considered as saponifiable lipids like waxes, triglycerides, sphingolipids, and phospholipids [3]. Further, lipids are categorized into polar and nonpolar lipids, or as simple lipids, complex lipids, and derived lipids (Fig. 24.1). A detailed resource of widely accepted lipid classification and nomenclature can be found at www.lipidmaps.org.

Many lipids show structural and functional similarities and dissimilarities across organisms. Although lipids are available in significant amount in mammalian organisms, the oleaginous (oil-bearing) microbial species can readily accumulate lipids [4–6]. Also known as single cell oils, these systems are considered as perfect tool to study advance lipidomics due to their rapid growth rates and their ability to grow on a wide range of substrates [7].

In recent years, lipid interests have emerged as therapeutic targets. Numerous studies have already been done and many more are in process to address their biological properties and significance. A specific example is the structure and function elucidation of complex glycosphingolipids. For example, glucosylceramide (GlcCer) biosynthetic pathways have been explored in detail to elaborate their roles in the regulation of fungal virulence and their specific roles in mammalian systems

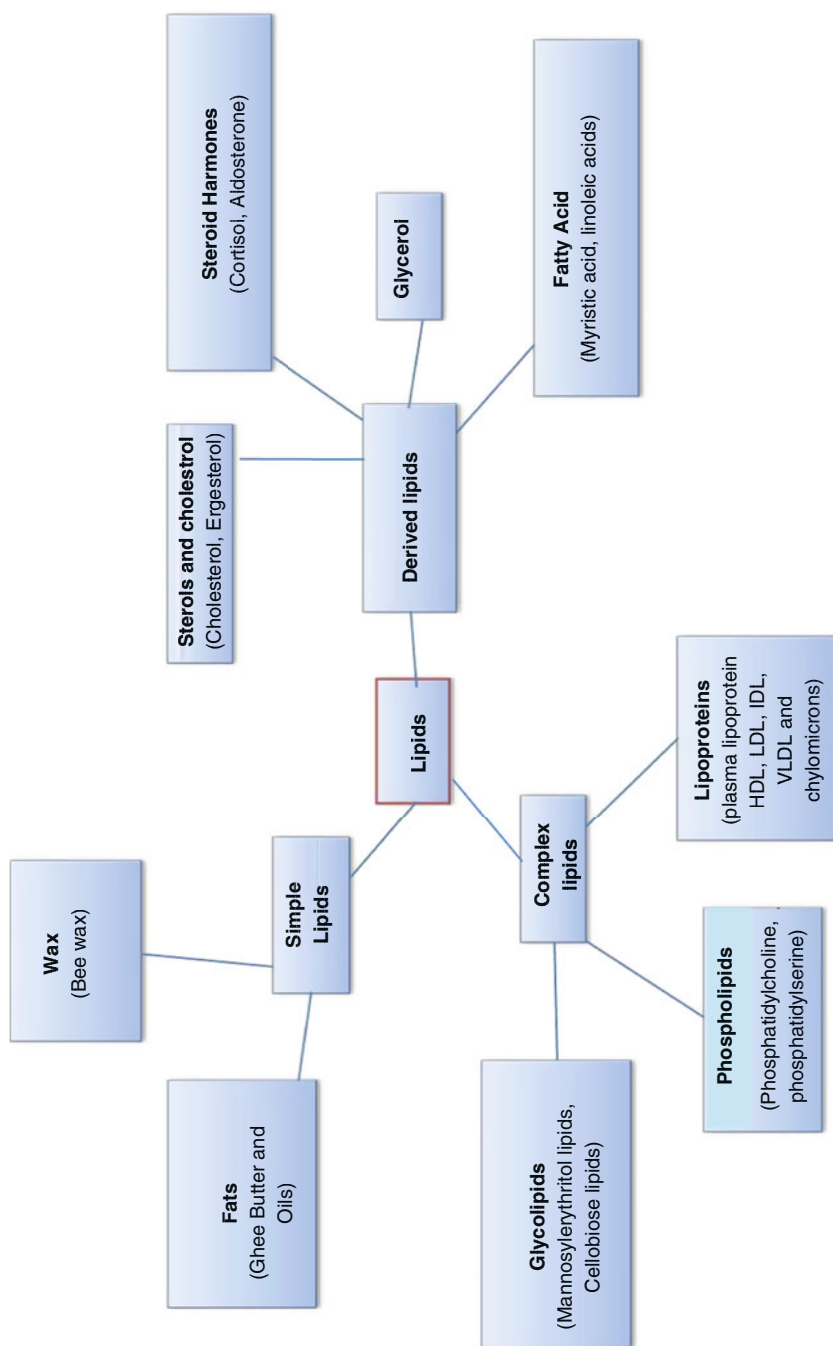


Fig. 24.1 Broad lipid classes and their examples

[8–10]. These studies open innovative venues for lipid-targeted antifungal therapeutics.

24.2 Lipid Classes and Structure

In biological organisms, lipids are essentially involved in membrane systems, in extracellular products, in cell wall synthesis, and also as storage material in abundantly observed lipid bodies. Diversity of lipid structures increases with cell size and complexity. Lipid composition and its type and concentration also vary with age, morphology and nutritional and environmental conditions [7, 11, 12]. In yeasts, it has been observed that in varying culture conditions *in vitro*, the lipid contents can be manipulated. Studies show a wide range of values for the contents of both polar and neutral lipids from the lipid fractions of different molds. Triacylglycerols are the major lipid component in the fungal body. It is considered as storage lipids and used as a carbon source for energy during growth and development. Various sterols, squalene, and hydrocarbons also majorly contribute their proportion in the lipid content of a fungus [13, 14].

Lipid molecules show hydrophobic and amphipathic nature. This can be attributed to it may be due to condensations of thioesters (carbanion-based) or by condensations of isoprene units (carbocation-based) like prenols and sterols. Lipid classes are vastly distributed based on structure and function that include fatty acids, hydrocarbons, glycerols, sterols and sterol esters, waxes, phospholipids, glycolipids, glyceride ethers, sphingolipids, and fat-soluble vitamins [15]. As shown in Fig. 24.1, these lipids are divided into three broad classes, as simple, complex, and derived lipids. Simple lipids are basically esters of fatty acids with different alcohols and divided into a subgroup of oil and fats. The hydrolysis of simple lipids yields a maximum of two primary products, whereas the hydrolysis of complex lipids yields three or more primary products per mole. Complex lipids usually contain a polar phosphate group and a glycerol backbone. Complex glycolipids contain a characteristic carbohydrate moiety [16]. Complex or compound lipids are divided into subcategories, for example:

1. Phospholipids that contain a phosphate moiety in addition to alcohol and fatty acid. They often contain nitrogenous and other substituent groups. In sphingophospholipids, the alcohol is a long-chain base sphingosine and in glycerophospholipids, glycerol is the alcohol [17, 18].
2. Glycolipids, such as glycosphingolipids, that are composed of a carbohydrate, a sphingosine, and a fatty acid [19].
3. Some other complex lipids like lipoproteins, sulfolipids, and amino lipids are also included in this category. When these simple and complex lipids work in a collaborated manner, they are known as derived lipid. Some key examples include hormones, ketone bodies, fat soluble vitamins, etc. [16].

Commonly occurring lipid structures consist of fatty acids linked by an ester bond to the glycerol or to amines or by amide bonds to sphingoid bases or to other alcohol groups [20]. Fatty acids are the carboxylic acids attached with aliphatic chain that are synthesized via condensation of malonyl-coenzyme A units by a fatty acid synthetase complex. Fatty acids are the reservoir of energy because, when metabolized, they produced large quantities of ATP. In nature, fatty acids generally have an 4 to 28 carbon atoms. In triacylglycerol lipids, three fatty acids are attached to a three-carbon property that came from two different kinds of lipids: it may be hard waxy solids at room temperature (fats) or translucent liquids (oils) [21, 22]. Fatty acid can be either unsaturated or saturated. Saturated fatty acids have no double bonds and possess higher melting points compared to the unsaturated fatty acids. They pack their molecules together and make straight rod-like shape [23, 24]. Unsaturated fatty acids are unbranched and contain *cis*-double bonds. A structural kink is created due to the presence of these *cis*-double bonds that disables the molecules to be grouped in straight rod-like shape [25].

Sterols are another key lipid class in biological systems. It is a wax-like substance, essential for the synthesis of cell membrane. In the cell membrane, the steroid ring structure provides hydrophobicity that boosts the rigidity and maintains the fluidity of the cell membrane. In animals, sterol is found as cholesterol that is synthesized in the liver and found in large concentrations within the spinal cord and brain. It is an organic compound containing a 3-hydroxylated cholestane core, containing a double bond at the 5,6-position. Cholesterol exists as a solid, neutral, and insoluble in water [26, 27]. In case of fungi, this role is played by ergosterol. Ergosterol is generally a phytosterol having an ergostane core that consists double bonds at the 22,23-, 5,6- and 7,8- positions and a 3 β -hydroxy group as well. It is a 3 β -sterol, an ergostanoid, and a member of phytosterols [28, 29].

Polyunsaturated fatty acids (PUFAs) contain multiple double bonds and exist as cellular lipids consisting of fatty acyl chains like phospholipids or triacylglycerols in vivo. Based on terminal double-bond location, it is classified into four groups, i.e., ω 9, ω 7, ω 6, and ω 3. Other major fungal lipids, based on their abundance, are palmitic acid (16:0), oleic acid (18:1), and linoleic acid (18:2) that are present in large quantities, while palmitoleic acid (16:1), linolenic acid (18:3), and stearic acid (18:0) are present in low concentrations. Further, unique lipid compositions have been reported in different fungal species in literature. It is documented that in mucorales, a unique γ -linolenic acid has been seen in its lipid composition rather than the docosahexaenoic acid (22: 5), eicosatetraenoic acid (20:4), and eicosapentaenoic acid (20:5) besides the occurrence of high amounts of PUFAs, C18:2, and C18:3.

Another lipid, steryl glucosides (SGs), is found in fungi while is rare in mammalian cells. These are the sugar derivatives of membrane-bound ergosterol in case of fungi and characterized by a planar sterol structure with four fused aliphatic rings and a hydrocarbon side chain at carbon 17 with the sugar moiety attached to the 3 β -hydroxy group at C3 of the sterol. In the second ring, a double bond is also attached in the backbone of sterol between C5 and C6 or between C7 and C8. Commonly occurring SGs in fungi are ergosterol-3 β -glucoside [30].

GlcCer is also majorly found in fungi. It is synthesized in the Golgi apparatus but primarily localized in the plasma membrane and also found at the budding sites of dividing cells in the cell wall and extracellular vesicles. Due to significant variation in the structure among fungal species, it is highly studied nowadays as new therapeutic target. GlcCer possesses β -linked glucose moiety and a ceramide backbone [30, 31]. A detailed classification of lipids based on their structure and functions is listed in Table 24.1.

24.3 Function of Lipids

Lipids play an essential role on energy storage and in formation of cell membranes. They are involved in diverse and widespread biological functions in the body in terms of intracellular signaling or local hormonal regulation, etc. Complex biosynthetic pathways are run to synthesize most of the lipid in the body. However, some essential lipids are supplemented through diet.

Different classes of lipid perform distinct roles. For example, fats and oils are used by the organism as an energy source, make a smooth safety layer of fat on to the skin, etc. They are involved in vital physiological functions for proper growth and development as well. A recent study showed the role of vegetable oil in growth of *A. melanogenum* [32]. It is documented that when vegetable oil is used in 1% with a carbon source, enhanced growth is recorded. It may be due to the presence of carbon-based long fatty acid chain compounds known as oleic acid, which are already used by several yeasts for growth like *S. cerevisiae*, *Candida* sp., *A. nidulans*, etc., in the vegetable oil [32, 33]. Glycerol is also required for growth in many fungal species as carbon and energy source like *A. nidulans*, *Candida* sp., *C. curvatus*, *S. cerevisiae*, *F. oxysporum*, etc. [34, 35].

Steroid hormones play crucial roles in metabolism regulation, inflammation, immune system, salt and water balance mechanism, and sexual development processes and also provide the ability to fight against injuries and diseases [36]. Cholesterol also participates in several enzymatic reactions. It is the major precursor for the synthesis of steroid hormones (like cortisol) and vitamin D. It actively participates in the brain synapses and the immune system. In humans, it is involved in bile acid biosynthesis, steroid biosynthesis, the lovastatin action pathway, and the zoledronate action pathway. It is also involved in several metabolic disorders like the child syndrome pathway, adrenal hyperplasia type 5, or congenital adrenal hyperplasia. Animal foods are the rich source of cholesterol. Different types of cholesterol required by the blood cells are low-density, high-density, and very low-density lipoproteins [37, 38].

Steroids play important roles in fungal biology and hence can be a novel target for antifungals, notably azoles. In humans, these are present in animal skin and get converted in ergocalciferol by ultraviolet rays. Ergosterol is the most abundant sterol structure in the fungal cell membrane and regulates membrane fluidity and function [39, 40]. Sterols also serve as precursors of steroid hormones in the fungal reproduction process. Phospholipids, commonly found in molds as a polar lipid, also include

Table 24.1 Classification of lipids based on its structure and function

Classification of lipids	Types of lipids	Composition	Synthesized/derivatives	Subtypes of lipids
Storage lipids (neutral) Simple lipids (organic compounds formed from alcohol + FA combined with ester linkage)	1) Fats 2) Wax	Ester of FA with glycerol Ester of higher FA with higher monohydric aliphatic alcohols (FA + a long-chain alcohol)	Triglycerides (naturally occurring oils and fats are a mixture of triglycerides)	Esters of trihydric alcohol, glycerol +3 FA
Membrane lipids (polar) Complex/compound lipids (ester of F.A, with alcohol)	1) Phospholipid (alcohol + phosphate+ FA) • Glycerophospholipid (alcohol is glycerol) • Sphingophospholipid (alcohol is sphingosine, amino alcohol instead of glycerol) 2) Glycolipids/ glycosphingolipids (glycerol+ FA + sugar) • Sphingolipids. • Galactolipids (sulfolipids/ sulphoglycosphingolipids	1) 2 FA + glycerol/other alcohol+ phosphoric acid, a nitrogenous containing base and other substrate 2) FA + sphingosine+ carbohydrate Sphingolipids: (FA + long-chain amino alcohol sphingosine via amide linkage)	Derivatives of phosphatidic acid. Amphipathic molecule containing a polar head and hydrophobic portion	Types of glycerophospholipid • Phosphatidylcholine (lecithin) • Phosphatidylethanolamine • Phosphatidylserine • Phosphatidylinositol • Phosphatidic acid • Cardiolipin • Plasmalogen • Platelet-activating factor • Phosphatidylglycerol Types of sphingophospholipid • Sphingomyelin Types of glycolipids • Cerebrosides. • Sulfatides • Globosides • Gangliosides Types of sphingolipids • Ceramide • Sphingomyelin • Neutral glycosides Glucosylcerebroside

(continued)

Table 24.1 (continued)

Classification of lipids	Types of lipids	Composition	Synthesized/derivatives	Subtypes of lipids
Derived lipids	1) Fatty acids (FA) • Saturated • Unsaturated Monounsaturated FA Polyunsaturated FA (eicosanoides: oxygenated derivatives of polyunsaturated FA) • Branched chain FA • Substitute FA	Aliphatic (long-chain) carboxylic acid	Eicosanoides: 20 carbon- containing FA generated from arachidonic acid Types of eicosanoides • Prostanoides • Prostaglandins • Thromboxanes • Prostacyclins • Leukotrienes • Lipoxins	• Gangliosides GM2 • Globosides Types of saturated FA • Lauric acid (C12) • Myristic acid (C14) • Palmitic acid (C16) • Stearic acid (C18) • Arachidic acid (C20) Types of unsaturated FA • Monounsaturated (monoethenoid, monoenoic) acids • Polyunsaturated (polyethenoid, polyenoic, linoleic acids, linolenic acids) acids • Oleic acid
	2) Glycerol (glycerin)	Trihydric alcohol (2-OH group)	Can be obtained from diet, from lipolysis of fats in adipose tissue, and from glycolysis	Can be utilized for the synthesis of triacylglycerols, phospholipids, and glucose or can be oxidized to provide energy
	3) Sterols and cholesterol	Steroid nucleus: four fused rings – hydroxyl group (polar head) in the A-ring – various nonpolar side chains	Synthesized in many tissue from acetyl-CoA and precursor of all steroid in the body (corticosteroid, sex hormone, bile acids, and vitamin D)	Types of steroid • Cholesterol • Testosterone • Estradiol • Ecdysone

<p>4) Steroid hormones</p>	<p>3 cyclohexanes + 1 cyclopentane fused together</p>	<p>Steroids derived from cholesterol (oxidized derivatives of sterols)</p>	<p>Types of steroid hormone</p> <ul style="list-style-type: none"> • Testosterone • Estradiol • Cortisol • Aldosterone • Prednisone • Prednisolone • Brassinolide • Stigmasterol (a plant sterol) • Ergosterol (a sterol from fungi and yeast)
<p>5) Biologically active lipids</p> <ul style="list-style-type: none"> • Lipid soluble vitamins (A, D, E, and K) • Arachidonic acid derivatives as signaling lipids • Bile salts (sodium cholate) 	<p>Isoprenoid derivatives</p>	<p>Acts as signaling molecules between nearby cells</p>	<p>Types of other isoprenoids</p> <ul style="list-style-type: none"> • Limonene • Bactoprenol • Juvenile hormone I

glycolipids sometimes and are essential components for the structure of biological membranes. They actively participate in the transportation of ions across membranes and activate some membrane-bound enzymes. In fungi, studies have shown that lipid compositions can drastically vary according to the age, culture conditions, and developmental stage. Lipid concentrations may also regulate the morphological factors.

Studies have shown that fungal lipids play crucial roles in virulence, defense by host, and the host-pathogen interaction. In recent years, interest in microbial lipids has been increased because of an urgent need of new therapeutic targets against drug-resistant microbes in medical and nutritional research. The potential use of PUFAs has been studied in this regard. To date, PUFAs have been produced commercially using microorganisms to treat several diseases like eczema, Alzheimer, cancer, depression, and peroxisomal disorders, but still the known data and uses about PUFAs are very limited [13]. PUFAs or their precursors are necessary for physiological activities; as mammals cannot synthesize all PUFAs, they are supplemented through diet. Mammals are also unable to desaturate oleic acid to linoleic acid, which is essential for cell growth and development. The secondary by-products of linoleic acid metabolism are crucial precursors for the synthesis of some hormones like prostaglandins, leukotrienes, and thromboxanes. Studies show that prostaglandins are an important hormonal substance for platelet aggregation, regulation of blood pressure, and immune system [13]. PUFAs also play critical roles in multiple aspects of membrane permeability, enzymatic activity, and signaling mechanism.

For the evaluation and comparison of the amount of microbial biomass, the signature fatty acids can be used due to its species-specific property. The concentration of different lipid classes at the different development stage and morphological character in same fungal species also show the lipid role in structural and functional regulation. [41] have shown this phenomenon in a study where they documented the role of lipid class and amount in the different stages of arbuscular mycorrhizal (AM) fungi. AM fungal spores and vesicles have a large amount of triacylglycerols as major neutral lipids. In contrast, neutral lipids rarely appeared in arbuscules. It is also seen that the total neutral lipid fatty acid concentration of the extraradical mycelia was several times higher in *S. calospora* than in *G. intraradices* [41].

In order to know the extended functions of other lipids, sugar derivatives of membrane-bound ergosterol like SGs have been established for their role in fungal virulence. Studies suggest that the *SGL1* deletion results in accumulation of SGs, which alter the fungal virulence properties [30]. Another lipid majorly studied for fungal virulence and defense mechanism is GlcCer. To analyze its role in pathogenicity, studies have been focused to delete the genes of respective enzymes involved in GlcCer biosynthesis or chemical manipulation of these enzymes. It has been seen that the disruption of GlcCer synthase in *C. neoformans* makes it avirulent and incompetent to grow at neutral/alkaline pH. Also, in *F. graminearum*, it has been seen that the depletion of GlcCer synthase shows a major alterations in conidia morphology and defective growth [8, 42].

24.4 Mass Spectrometry-Based Yeast Lipidomics

Over the last 25 years of using mass spectrometry technology leads to a new way of identifying and quantifying macromolecules in a high-throughput manner. Mass spectrometry is an analytical technique capable of identification and quantification of small molecules in a sample. It also helps in determination of unknown compounds by elucidating their molecular structure. Earlier lipids were thought to function only for storage and structural purposes, but in recent years they are discovered to be metabolites with versatile functions owing to its versatility to interactive and dynamic nature. Compared to the transcriptomic and proteomic studies, the lipidomic study did not gain wide recognition because of the difficulty in analyzing and quantifying the full lipidome of the organism. No such methods and techniques were known to fully characterize every lipid molecular species. Traditional methods of lipid analysis include gas chromatography, thin-layer chromatography, and metabolic labeling. These techniques separate the lipids based on only mass, are of low resolution, and are not selective and sensitive as compared to the present-day mass spectrometry techniques [43]. Recently there is a rapid advancement in the field of lipidomics due to the advancement in the technology such as high-resolution mass spectrometry and specific lipid extraction protocols. Development of soft ionization techniques and specially designed computational software has allowed specific and selective detection.

In yeasts, there are three major classes of lipids, namely, glycerophospholipids, sphingolipids, and sterols. They all differ from each other in their core structure or head group and fatty acyl chain. Glycerophospholipid heads are conserved in different eukaryotes. Sphingolipid composition is relatively distinct than mammals. The major sphingolipids are complex phytoceramide derivatives or hexose sugar-linked ceramide structures. Ergosterol is the major sterol structure present [43, 44]. In 2009, the first and most detailed description of the lipidome of the *S. cerevisiae* was published [44]. Using the quantitative shotgun mass spectrometry approach, they quantified 250 lipid molecular species, classified into 21 classes including inositol-containing sphingolipids, intermediate lysophospholipids, and bioactive long-chain bases which were not identified before. They used a two-step lipid extraction procedure and separated the polar and nonpolar lipids and optimized the solvent system which resulted in improved ionization efficiency in negative ion mode and detection sensitivity for the anionic lipid classes to be used in the mass spectrometric analysis. The most abundant lipid species identified ergosterol, followed by mannosyldiinositolphosphorylceramide, and other sphingolipids intermediates like mannosylinositol phosphorylceramide, ceramide, and inositolphosphorylceramide. Glycerophospholipid species contained 16:0, 16:1, and 18:1 fatty acid moieties. They also found that at different temperature conditions, compositions of lipid molecular species change. Furthermore, they analyzed the change in the composition of lipids in $\Delta elo1$, $\Delta elo2$, and $\Delta elo3$ mutants. The content of inositolphosphorylceramide derivatives in $\Delta elo1$ mutant was high. On the other hand, the levels of phosphatidylinositol, phosphatidylcholine, and phosphatidylethanolamine were reduced in $\Delta elo1$ mutant. The lipidome of

$\Delta elo2$ and $\Delta elo3$ mutants showed specific changes in the molecular lipid species compositions. They also observed an increase in long-chain base, and sphingolipidome showed less mannosylinositolphosphorylceramide and mannosyldiinositolphosphorylceramide but more inositolphosphorylceramide. They observed that in both $\Delta elo2$ and $\Delta elo3$ mutants, the molecular compositions of phosphatidylinositol and phosphatidylcholine were different. The information revealed by this lipidomic study laid out the foundation stone for the currently flourishing field of high-throughput lipidomics.

In recent times various chromatographic techniques are coupled with mass spectrometric techniques to analyze the structural as well as functional lipids in different species at different conditions. Such platforms have been routinely used to analyze the lipidomic variations of yeasts like *S. cerevisiae* and *S. pombe* [45]. The phospholipid and triacylglyceride composition of *S. pombe* contains 18:1 fatty acyls and that of *S. cerevisiae* contains 16:0, 16:1, and 18:1 fatty acid moieties. *S. cerevisiae* has high levels of inositolphosphorylceramide, while *S. pombe* cells are enriched in phytoceramides.

24.5 Triple Quadrupole Mass Spectrometry Approach for Lipidomics

A triple quadrupole mass spectrometry setup is quite useful for large-scale targeted analysis of lipid intermediates. Triple quadrupole mass spectrometry has been used to analyze lipid biomarker candidates from plasma for Alzheimer's disease [46]. Using a multiple reaction monitoring approach, researchers could perform a quantitative analysis of targeted lipid molecular species of 22 classes and determine a comprehensive lipid profile that provided intriguing insights into the lipid biomarkers of Alzheimer patients. This technique also helps in capturing changes of the individual lipid species and lipid molecular species. Another study compared the triple quadrupole with quadrupole time of flight (Q-TOF) mass spectrometry for phosphopeptides analysis using the precursor ion scanning in both negative and positive nano-electrospray (nESI) ionization modes [47]. The study showed that selectivity and sensitivity of precursor ions for a particular mass-to-charge ratio in Q-TOF are better than the triple quadrupole because of its high resolution. The sensitivity of the Q-TOF system was significantly better because the collision Q2 quadrupole enhances the transmission of the select m/z ions from the collision cell into the TOF analyzer. In some classical studies, the structure of phosphatidylglycerol and phosphatidylethanolamine could be identified earlier by the triple quadrupole tandem mass spectrometry approach using electrospray ionization [48, 49]. In a more recent study, the oxidized membrane lipids could be detected using the liquid chromatography coupled with mass spectrometry, with high selectivity and sensitivity [50]. Below we describe some technical aspects of quadrupole mass analyzer.

24.5.1 Quadrupole

A quadrupole is a highly selective and sensitive setup to filter out specific m/z ions and has been extremely useful in routine lipidomic setups over the years. A quadrupole is a mass analyzer, which consists of four rods that are arranged parallelly. These rods are electrodes with DC and RF voltages. If only the RF voltage is applied, many ions of different m/z will pass through the quadrupole. On the other hand, when a combination of RF and DC voltage is applied to the rod, molecules of specific m/z ratio will pass through the rods based on the magnitude of DC and RF. When DC and RF voltages are increased and their ratio remains constant, the mass range can be scanned to acquire the mass spectrum. Each pair of rods is connected, and the rods have the same voltage as the opposite rods. Two rods have $+V_{DC}$ and $+V_{RF}$, and the other two rods have negative $-V_{DC}$ and $-V_{RF}$. For a certain period, both the RF and DC fields increase. The amplitude of the RF field is stronger than the DC fields. There are two requirements for the ions to pass through the rods:-

1. The time required for crossing the analyzer should be short.
2. The ions should remain for sufficient time between the rods so that few oscillations of the alternative occur.

Quadrupole focuses the ion at the center of the rods; if the positive ion reaches near the positive rod, its potential energy increases, and decreases near the negative rods. Alternative field spins the potentials.

24.5.2 Triple Quadrupole

The triple quadrupole mass analyzer is a tandem mass spectrometry method. It is used for knowing the structural information and quantification of the molecules. Triple quadrupole consists of three rods parallel to each other. Quadrupole 1 (Q1) and quadrupole 3 (Q3) act as mass filters. Quadrupole 2 (q2) is a collision rod, which causes fragmentation of the analyte by collision with a gas. The q2 is a radiofrequency-only quadrupole. Together these are able to detect the molecular structure and quantity in SRM and MRM mode of any targeted analyte. Quadrupole can be used for scanning and filtering. During scan mode, DC and RF voltages were increased resulting in the acquisition of full scan mass spectra but at the expense of low sensitivity and low speed. First, the Q1 is set to scan the analyte of a specific m/z ratio. q2 and Q3 can pass all ions. These quadrupoles can be used in various ways to provide the information about specific select ion(s) [50].

A triple quadrupole is designed to work in four modes: (i) precursor ion scan: is used for identifying a class of compounds which are having common daughter ion. In Q1 scanning takes place and Q3 is set for daughter ion which is generated in q2 by collision-induced dissociation (CID). (ii) Neutral loss scan: unchanged fragments are identified by scanning a neutral mass between Q1 and Q3. (iii) Product ion scan: this

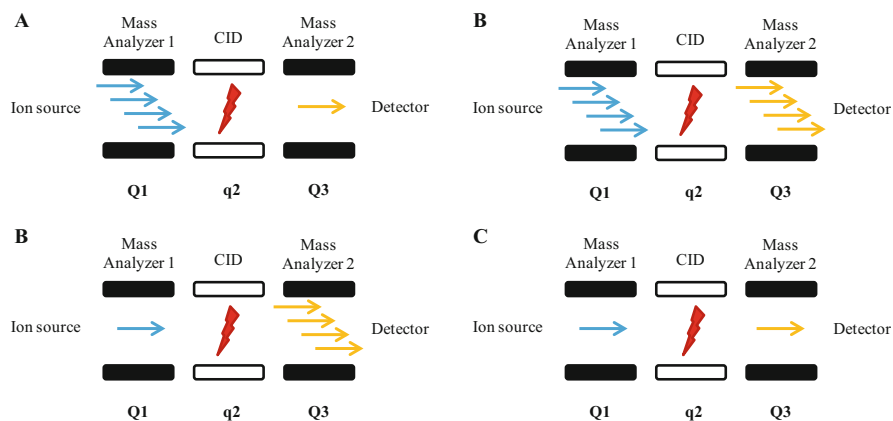


Fig. 24.2 Representative scheme of a triple quadrupole mass spectrometer. Ions generated by soft ionization techniques like electrospray ionization can be detected by the quadrupoles. (a) Precursor ion scan: the molecular ions are detected in Q1, followed by CID in q2, and the select daughter ions are detected in Q3. For example, phosphatidylcholine molecular species can be detected using the positive precursor scan of m/z 184. (b) Neutral loss scan: a defined mass difference in parent and daughter ions are recorded in Q1 and Q3, respectively. For example, loss of a water molecule results in a mass difference of 18 Da. (c) Product ion scan: a select ion in Q1 undergoes dissociation in q2, and its spectra are recorded in Q3. (d) Single reaction monitoring: specific m/z ions are detected in Q1 and Q3, dissociation occurs at q2. For example, GlcCer(d19:2/18:0(2OH) molecular species is detected using the m/z 756 \rightarrow m/z 276 reaction in a triple quadrupole mass spectrometer. It must be noted that optimization voltages and collision energy among several other parameters are crucial for accurate detection. The figure has been modified from [51]

mode gives the structural information about the specific parent ion, formed in the Q1, then in q2 fragments are generated by CID further they are scanned in Q3. (iv) Single or multiple reaction monitoring: these are performed by the selection of a parent ion in Q1, fragmentation from q2, and analyzing the characteristics of fragments in Q3 (see Fig. 24.2).

In triple quadrupole, ion is filtered twice without interference from the large number other compounds present in the extracellular matrix. The triple quadrupole mass spectrometry has wide-scale applications for the quantitative analysis of lipid molecular species. There are certain advantages of triple quadrupole over single quadrupole. Triple quadrupoles are fast, robust, time-efficient, and cost-effective compared to single quadrupole. They are more selective and provide high sensitivity during lipidomic applications.

In conclusion, this chapter describes key aspects of lipid classification, function, structure, and lipidomic applications. The lipid analysis represents a major aspect of lipid research in cellular organisms. Here we have described the triple quadrupole mass spectrometry-based analysis of lipids as a significant technique for lipidomic studies.

Acknowledgments This work was supported by grants from ICMR (No.52/08/2019-BIO/BMS) to AS/SM, DST-PURSE program (SR/PURSE Phase 2/29(C)) to AS/SM, DBT (BT/PR38505/

MED/29/1513/2020) to AS, ICMR grant (No. 56/2/Hae/BMS) to NB, and COE Govt. of UP grant (No. 66/2019/1864/Sattar-4-2019-4(24)/2019) to SM and (No. 10/2021/281/सत्तर-4-2021-04(2)/2021) to AS. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

References

1. Gurr MI, Harwood JL, Frayn KN (2002) Lipid biochemistry, vol 409. Blackwell Science, Oxford
2. Konige M, Wang H, Sztalryd C (2014) Role of adipose specific lipid droplet proteins in maintaining whole body energy homeostasis. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1842(3):393–401
3. Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CR, Shimizu T, Dennis EA (2009) Update of the LIPID MAPS comprehensive classification system for lipids. *J Lipid Res* 50 (Supplement):S9–S14
4. Papanikolaou S, Aggelis G (2011a) Lipids of oleaginous yeasts. Part I: biochemistry of single cell oil production. *Eur J Lipid Sci Technol* 113(8):1031–1051
5. Papanikolaou S, Aggelis G (2011b) Lipids of oleaginous yeasts. Part II: technology and potential applications. *Eur J Lipid Sci Technol* 113(8):1052–1073
6. Ratledge C, Wilkinson SG (1988) Microbial lipids, vol 2. Academic Pr
7. Athenaki M, Gardeli C, Diamantopoulou P, Tchakouteu SS, Sarris D, Philippoussis A, Papanikolaou S (2018) Lipids from yeasts and fungi: physiology, production and analytical considerations. *J Appl Microbiol* 124(2):336–367
8. Munshi MA, Gardin JM, Singh A, Luberto C, Rieger R, Bouklatas T, Del Poeta M (2018) The role of ceramide synthases in the pathogenicity of *Cryptococcus neoformans*. *Cell Rep* 22 (6):1392–1400
9. Takahashi T, Nieda M, Koezuka Y, Nicol A, Porcelli SA, Ishikawa Y, Juji T (2000) Analysis of human α 24+ CD4+ NKT cells activated by α -glycosylceramide-pulsed monocyte-derived dendritic cells. *J Immunol* 164(9):4458–4464
10. Tsuji M, Gonzalez-Aseguinolaza G, Koezuka Y (2009) U.S. patent no. 7,488,491. U.S. Patent and Trademark Office, Washington, DC
11. Murphy DJ (1993) Structure, function and biogenesis of storage lipid bodies and oleosins in plants. *Prog Lipid Res* 32(3):247–280
12. Weete JD (2012) Lipid biochemistry of fungi and other organisms. Springer Science & Business Media
13. Akpınar-Bayazit A (2014) Fungal lipids: the biochemistry of lipid accumulation. *Int J Chem Engg App* 5(5):409
14. Beopoulos A, Nicaud JM, Gaillardin C (2011) An overview of lipid metabolism in yeasts and its impact on biotechnological processes. *Appl Microbiol Biotechnol* 90(4):1193–1206
15. Bose, T. K., & Sinha, P. K. (2009). *Biomolecules*
16. Fahy E, Subramaniam S, Brown HA, Glass CK, Merrill AH Jr, Murphy RC, Shimizu T (2005) A comprehensive classification system for lipids. *Eur J Lipid Sci Technol* 107(5):337–364
17. Lordan R, Tsoupras A, Zabetakis I (2017) Phospholipids of animal and marine origin: structure, function, and anti-inflammatory properties. *Molecules* 22(11):1964
18. Miñones J Jr, Pais S, Miñones J, Conde O, Dynarowicz-Łątka P (2009) Interactions between membrane sterols and phospholipids in model mammalian and fungi cellular membranes—a Langmuir monolayer study. *Biophys Chem* 140(1–3):69–77
19. Kates M (1990) Glycolipids of higher plants, algae, yeasts, and fungi. In: *Glycolipids, Phosphoglycolipids, and Sulfoglycolipids*. Springer, Boston, MA, pp 235–320

20. Arora A, Gupta CM (1997) Glycerol backbone conformation in phosphatidylcholines is primarily determined by the intramolecular stacking of the vicinally arranged acyl chains. *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1324(1):47–60
21. Rusta AC, Drevon CA (2001) Fatty acids: structures and properties. e LS
22. Stahl PD, Klug MJ (1996) Characterization and differentiation of filamentous fungi based on fatty acid composition. *Appl Environ Microbiol* 62(11):4136–4146
23. Brenner RR (1984) Effect of unsaturated acids on membrane structure and enzyme kinetics. *Prog Lipid Res* 23(2):69–96
24. Mayes PA, Botham KM (2003) Metabolism of unsaturated fatty acids and eicosanoids. a LANGE medical book, 190
25. Rioux V, Legrand P (2007) Saturated fatty acids: simple molecular structures with complex cellular functions. *Current Opinion in Clinical Nutrition & Metabolic Care* 10(6):752–758
26. Fasman GD (ed) (2019) Handbook of biochemistry and molecular biology. CRC Press, Lipids Carbohydrates, Steroids
27. Myant NB (2014) The biology of cholesterol and related sterols. Butterworth-Heinemann
28. Parks LW, Casey WM (2017) Fungal sterols. In: Lipids of pathogenic fungi (1996). CRC Press, pp 63–82
29. Rodrigues ML (2018) The multifunctional fungal ergosterol. *MBio* 9(5)
30. Rella A, Farnoud AM, Del Poeta M (2016) Plasma membrane lipids and their role in fungal virulence. *Prog Lipid Res* 61:63–72
31. Athanasopoulos A, André B, Sophianopoulou V, Gournas C (2019) Fungal plasma membrane domains. *FEMS Microbiol Rev* 43(6):642–673
32. van Nieuwenhuijzen EJ, Sailer MF, van den Heuvel ER, Rensink S, Adan OC, Samson RA (2019) Vegetable oils as carbon and energy source for *Aureobasidium melanogenum* in batch cultivation. *Microbiol Open* 8(6):e00764
33. Del Rio JL, Serra P, Valero F, Poch M, Sola C (1990) Reaction scheme of lipase production by *Candida rugosa* growing on olive oil. *Biotechnol Lett* 12(11):835–838
34. Castro IM, Loureiro-Dias MC (1991) Glycerol utilization in *Fusarium oxysporum* var. lini: regulation of transport and metabolism. *Microbiology* 137(7):1497–1502
35. Chatzifragkou A, Makri A, Belka A, Bellou S, Mavrou M, Mastoridou M, Papanikolaou S (2011) Biotechnological conversions of biodiesel derived waste glycerol by yeast and fungal species. *Energy* 36(2):1097–1108
36. Yadav R, Yadav N, Kharya MD (2014) Steroid chemistry and steroid hormone action: a review. *Asian J Res Chem* 7(11):964–969
37. Bloch K (1991) Cholesterol: evolution of structure and function. In: New comprehensive biochemistry, vol 20. Elsevier, pp 363–381
38. Pfrieger FW (2003) Role of cholesterol in synapse formation and function. *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1610(2):271–280
39. Iwaki T, Iefuji H, Hiraga Y, Hosomi A, Morita T, Giga-Hama Y, Takegawa K (2008) Multiple functions of ergosterol in the fission yeast *Schizosaccharomyces pombe*. *Microbiology* 154(3):830–841
40. Zhang YQ, Gamarra S, Garcia-Effron G, Park S, Perlin DS, Rao R (2010) Requirement for ergosterol in V-ATPase function underlies antifungal activity of azole drugs. *PLoS Pathog* 6(6): e1000939
41. van Aarle IM, Olsson PA (2003) Fungal lipid accumulation and development of mycelial structures by two arbuscular mycorrhizal fungi. *Appl Environ Microbiol* 69(11):6762–6767
42. Ramamoorthy V, Cahoon EB, Thokala M, Kaur J, Li J, Shah DM (2009) Sphingolipid C-9 methyltransferases are important for growth and virulence but not for sensitivity to antifungal plant defensins in *Fusarium graminearum*. *Eukaryot Cell* 8(2):217–229
43. Guan XL, Riezman I, Wenk MR, Riezman R (2010) Yeast lipid analysis and quantification by mass spectrometry. *Methods Enzymol* 470:369–391

44. Ejsing CS, Sampaio JL, Surendranath V, Duchoslav E, Ekroos K, Klemm RW, Simons K, Shevchenko A (2009) Global analysis of the yeast Lipidome by quantitative shotgun mass spectrometry. *Proc Natl Acad Sci U S A* 106(7):2136–2141
45. Shui G, Guan XL, Low CP, Chua GH, Goh JS, Yang H, Wenk MR (2010) Toward one step analysis of cellular lipidomes using liquid chromatography coupled with mass spectrometry: application to *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* lipidomics. *Mol BioSyst* 6(6):1008–1017
46. Tokuoka SM, Kita Y, Shimizu T, Oda Y (2019) Isobaric mass tagging and triple quadrupole mass spectrometry to determine lipid biomarker candidates for Alzheimer's disease. *PLoS One* 14(12):e0226073
47. Steen H, Küster B, Mann M (2001) Quadrupole time-of-flight versus triple-quadrupole mass spectrometry for the determination of phosphopeptides by precursor ion scanning. *J Mass Spectrom* 36(7):782–790
48. Hsu FF, Turk J (2000) Characterization of Phosphatidylethanolamine as a Lithiated adduct by triple Quadrupole tandem mass spectrometry with electrospray ionization. *J Mass Spectrom* 35(5):595–606
49. Hsu FF, Turk J (2001) Studies on Phosphatidylglycerol with triple Quadrupole tandem mass spectrometry with electrospray ionization: fragmentation processes and structural characterization. *J Am Soc Mass Spectr* 12(9):1036–1043
50. Li L, Zhong S, Shen X, Li Q, Xu W, Tao Y, Yin H (2019) Recent development on liquid chromatography-mass spectrometry analysis of oxidized lipids. *Free Radic Biol Med* 144:16–34
51. Singh A, Del Poeta M (2016) Sphingolipidomics: an important mechanistic tool for studying fungal pathogens. *Front Microbiol* 14(7):501

Part V

**Translational Omics: Omics Applications in
Translational Research**



Nanoparticles as Therapeutic Nanocargos Affecting Epigenome of Microbial Biofilms

25

Indu Singh, Pradeep Kumar, and Gagan Dhawan

Abstract

Biofilm formation and its associated infections have been inflating as a serious health problem worldwide. Environmental factors that might affect biofilm formation are temperature, surface geometry, pH and oxygen content. Enzymatic action, binding proteins, polysaccharide adhesive protein, lipid content and quorum sensing encoding genes support biofilm to survive for a longer period. Adherence of bacteria to particular area restricts resource availability after a certain period that potentiates bacteria for dispersal and move towards the newer template. The extent of bacterial adhesion to the surface decides weak or strong biofilm formation. Regulation of biofilm-associated cellular functioning depends on different types of inducer (AI-1, AI-2) and lactone (AHL) moieties present beneath biofilm dense matrix. In the context of regulating cellular functioning, different types of inducer (AI-1, AI-2) and lactone (AHL) moieties are available beneath the biofilm, under the dense matrix. DNase, DNA methylation and small non-coding RNA played a key role in the survival as well as the dispersal process. DNA methylation is mandatory for performing replications. DNase is the dispersal enzyme meant for hydrolysing the extracellular DNA

I. Singh

Department of Biomedical Science, Acharya Narendra Dev College, University of Delhi, New Delhi, India

CSIR-Institute of Genomics and Integrative Biology, Delhi, India

P. Kumar

CSIR-Institute of Genomics and Integrative Biology, Delhi, India

G. Dhawan (✉)

Department of Biomedical Science, Acharya Narendra Dev College, University of Delhi, New Delhi, India

e-mail: gagandhawan@andc.du.ac.in

(eDNA) present in the cellular matrix of bacteria biofilm. Apart from understanding biofilm-forming mechanism, biofilm-associated infections and their treatment strategies have been explored using a nanobiotechnological approach such as the use of nanophages loaded with metal NPs, nanotubes fabricated with antibacterial drugs, nanorods, nanocapsules adsorbed with antimicrobial peptides (AMPs), AMPs loaded with metal NPs on hydrogels and nanozymes for treating medical device-associated pathological conditions. Simultaneously, epigenetic aspects were also discussed for inculcating nanostructures as epigenomic markers for detecting virulence factors of biofilm-forming bacterial isolates.

Keywords

Biofilm · Epigenetics · Antimicrobial peptides (AMPs) · Nanophages · Biofilm-associated infections

25.1 Introduction

Biofilm formation is a common phenomenon by single microbial species or multi-species on natural as well as artificially contemplated surfaces. Surface adherence, maturation and dispersion of participating microorganisms are a progressive multi-stage process for biofilm formation [1]. Environmental factors (such as temperature, surface composition, pH, oxygen level for aerobic microbes) enhance the formation of biofilms and its stability. Essential components of biofilm comprise polysaccharide intercellular adhesion (PIA), lipid, proteins, nucleic acid and extracellular genomic DNA (eDNA). Different microbes are involved in biofilm formation steps via an assorted pathway. The organic scaffold of the extracellular matrix contains mineral layer, mainly calcium carbonate, and other vital elements required for biofilm existence. Bacterial genetic environment triggers biomineralization beneath the film and sets up a unique framework for bacterial colonies that may protect them from antibiotic and environmental shock and allow a fair exchange of nutrients through ridges and fissures constructed within the biofilm. Secondary messengers like cAMP and C-di-GMP are the bridge forming factor between environmental factors and gene regulations in the biofilm, thus helping in the shaping of biofilm [2].

Genotypic and protein participation promote weak and robust biofilm formation. In the case of *Staphylococcus aureus* (SA), *agr* dysfunctioning leads to biofilm production. β -Haemolysin (toxin) is strongly associated with weak biofilm phenotype, while α toxin is encoded by *hla* gene, which strongly supports colonization on surfaces, thus forming a strong biofilm of *S. aureus* strain [3]. The imbalance between the genes encoding for binding proteins attribute for forming a weak and robust film. The extent of adhesion on the template surface describes the binding forces between the bacterial cell and surface. Various groups highlighted the bunch of genes (*pap*, *sfa*, *afa* and *LasB*, *rhlA*, *int1*, *int2*) involved in robust biofilm

production especially in gram-negative strains such as *Escherichia coli* [4, 5] and *Pseudomonas aeruginosa* [6], respectively.

Various nanoforms have been used as epigenomic markers for detection of virulence factors that give rise to biofilm-associated infection, which is a significant threat around the globe. Pathogenicity of bacterial cell leads to the development of resistance against traditional antibiotics. Antibiotic resistance motivates for designing numerous nanostructures with specific target delivery despite causing any side effects, non-immunogenic and without cytotoxicity. In this chapter, we discussed different aspects of biofilm formation, an enzymatic approach from initial development to dispersal progression. We also mentioned antimicrobial peptides (AMPs) loaded with antibacterial drugs or hydrogels, nanophages, nanocapsules, nanorods, nanotubes fabricated with antibacterial NPs and conjugated metal oxide NPs for antibiofilm potential. Epigenetic mechanism played a crucial role in the identification of virulence of bacterial biofilm. The epigenomic study is newly endorsed area for enumerating genes involved from initial adhesion to final maturation as well as dispersal process. Bacteria adhere to using diverse binding proteins and gene regulating these mechanisms. Epigenomic markers elicited for biofilm detection and treatment have been discussed in the present chapter.

25.2 Biofilm Formation and Its Potential Promoting Factors

25.2.1 Gene Regulation in Weak and Strong Biofilm

Every strain regulates its multiplication and exponential growth with the help of upregulation/downregulation of gene involved as well as proteins. Imbalance among the genes encoding for binding factors reveals the extent of bacterial strain adherence on the surface, which could be characterized as (1) weak biofilm and (2) robust biofilm. Herein, we discuss how gene regulates the weak and robust formation of MRSA. MRSA adhere to surfaces via binding proteins such as laminin, elastin and fibrinogen encoded by specific gene *eno*, *ebps* and *fib*, respectively. In a strong biofilm, levels of *ebps* gene were higher as compared to weak biofilm. While *eno* gene meant surface receptors, *fib* gene meant for adherence of SA to fibrinogen protein adsorbed on an indwelling catheter. At the initial phase of biofilm formation, these genes were upregulated that give rise to robust biofilm while slightly downregulated in weak biofilm [7]. Abad et al. stated that *pap*, *sfa* and *afa* adhesion-encoding genes are involved in the information of weak adhering and firm adhesion of *E. coli* strain on the surface [5]. Karami et al. reported numerous highly prevalent genes *LasB*, *rhlA*, *Int1* and *Int2* in *chronological* for strong biofilm-forming ability. *LasB* is the initiation gene that helps bacteria to initiate biofilm formation, while *rhlA* (quorum sensing regulon; QSR) attribute in the cellular signalling among bacterial cell. An identified group of genes and integrons enclaves identification of antibiotic resistance extremity in *P. aeruginosa* isolates [6].

25.2.2 Enzymatic Action on Biofilm Growth and Dispersal Process

Enzymes help in exopolysaccharide secretions, quorum sensing, biofilm maturation and dispersion [8]. Bacterial species secrete extracellular enzymes that enhance pathogenicity as well as virulence in the host. Predominantly *staphylococci* species secrete prominent families of enzymes such as proteases and lipases that aggravate the biofilm formation [9]. The virulent factor secreted by *Pseudomonas* species that involve LipC, LipA, EstA and LasB that plays a key role in establishing biofilm architecture. Oversecretion of LipC allows cell adhesion to each other, thus helping to form assorted biofilms. On the other hand, EstA and LasB overproduction by strains would not be able to form biofilm scaffold, as well as these enzymes act on bacterial cell motility via rhamnolipids release mechanism [10]. Extracellular enzymes are termed as virulent factors in almost all bacterial strains and attack on a substrate which ultimately causes bacterial attachment and empowers skin formation [11].

Biofilm dispersion is the major process in which bacteria from macro colonies escape with the help of self-generating biomolecules (autoinducers) which specifically target gene regulating quorum sensing. Usually, biofilm survives with the help of balance between EPS forming proteins and dispersing enzymes, but imbalance (upregulation of matrix-forming protein and downregulation of dispersing enzymes) leads to a chemical imbalance that triggers dispersal process immediately [12].

Besides, the following are the different forms of small molecules that are secreted which hasten the dispersal process:

- (a) Autoinducing peptides (AIPs) are secreted especially by gram-positive bacterial strains. Staphylococci species perform *Agr* system via upregulated expressions of peptidases/nucleases that hasten dispersal. Prominently, AIP-I is secreted by *methicillin-resistant Staphylococcus aureus* (MRSA) that augment MRSA absconding out of scaffold [13].
- (b) Autoinducers in gram-negative strains are secreted as homoserine lactones (HSLs); these HSLs bind to LuxR cytoplasmic protein (receptor) which enables QS transcription and targets QS signalling and DNA dimerization. When homoserine and lactone linked together with amide linkage and form acyl homoserine lactones (AHLs), these AHLs produce signals that promote biofilm dispersal process in gram-negative strains [14].
- (c) Another important signalling molecule is autoinducer-2 (AI-2; furanosyl borate diester) which is produced during conversion of homocysteine to S-ribosyl-homocysteine (SRH) via S-adenosyl-homocysteine (SAH) and secreted by both gram-positive and gram-negative strains. This molecule is proved to be responsible for inter- and intra-species communication (multispecies biofilm formation), cellular motility as well as dispersal [15]. AI-2 signalling has been expressed by *LuxS* gene which might be responsible for forming and maintaining biofilm templates [16].
- (d) Matrix-degrading enzymes, i.e. proteases [17], glycosidase [18], deoxyribonucleases [19], dispersin B [20] and hyaluronidase for *Streptococcus*

intermedius and *Staphylococcus aureus* [17], attribute for EPS disrupter and hamper sedentary lifestyle of bacterial cells beneath biofilm.

25.2.3 Epigenetic Makeup of Bacterial Population Beneath Biofilm and Cellular Signalling

Bacterial population beneath biofilm utilizes multiple factors that contribute to cellular signalling as well as metabolic and transcriptional processes. The most prominent factor is C-di-GMP that participates in the biofilm survival via tracking bacterial transcription, adhesive protein action, extracellular DNA (eDNA) secretion, cell density/cell motility and finally cell death. It determines the differences between floating bacteria and bacteria residing in biofilm [21]. Another factor is small non-coding RNA molecule which is a post-transcriptional regulatory molecule in the bacterial genome that determines the metabolic process, stress tolerance and virulence in a host such as *SrB* in *Pseudomonas aeruginosa*. It has been addressed that the foremost type of sRNA has been identified, i.e. trans-encoding sRNA post-transcriptional regulator contains a large number of binding sites for mRNA which ultimately assure broad transcriptional regulation as well as agonistic/antagonistic effect on translation process in the bacterial cell. The difference in the number of binding sites might be due to the number of nucleotides present in both sRNA and mRNA [22]. Binding of sRNA-mRNA leads to degradation of both molecules, hence uplifting pathogenicity in response to environmental cues. mRNA degradation and its short half-life regulate protein synthesis as well as allow adaption of bacterial genetic profile response for environmental stress conditions, thus aggravating virulence [23]. Sigma factor (σ) known as multi-domain subunit of bacterial RNA polymerase helps in RNA synthesis via different pathways. This sigma unit bind to another subunit ($\beta\beta'\alpha_2\omega$), therefore forming “holoenzyme” which helps to recognize promoter site and initiates steps for RNA synthesis that is a major part of biofilm formation. σ factor in *Bacillus subtilis* deals with stress responses as well as spore formation and thermal shock-responsive behaviour in *E. coli* [24]. Epigenetic study reveals the chemical compound that modifies gene expression depicted by bacterial community such as DNA methylation, micro non-coding RNA-based mechanisms, histone modifications, Dam and CcrM. The Dam (DNA adenine methyltransferase) enzyme in *E. coli* is essential for expression for a gene like *pap* operon [25].

Zinc ion (Zn^{2+}) is the known cofactor used by various metalloenzymes in bacterial metabolic activities. Zn^{2+} help in biofilm formation via promoting cell-to-cell adhesion, cell surface smoothening and antibiotic resistance, preventing it from oxidative stress and maintaining the integrity of bacterial cell wall [26]. Amount of Zn^{2+} plays a crucial role in bacterial survival; high concentration imposes toxic effects, but at low concentration, it helps in maintaining bacterial homeostasis. Regulation of zinc ion involves the participation of metaloregulatory protein, zinc uptake regulator (*Zur*)/transporters (ABC transporter) and efflux pumps (p-type ATPase, CDF family, RND family). It all depends on feedback mechanism, i.e. at normal concentration, Zn^{2+} uptake ABC transporter represses, but decreased

concentration activates ABC transporter to grab Zn^{2+} from the environment [27]. Zinc ion promotes virulence in host, thus stimulating antibiotic resistance mechanism.

25.2.4 Factors Affecting Genomic Environment of Microbial Biofilm

Biofilm comprised of disparate gene and proteins that participates in the complexity of the developmental process. Transcriptional *lacZ* reporter genes, *slp* and *OmpC* are the major genes in gram-negative bacterial strain (*Escherichia coli*) especially involved in the initial phase of biofilm formation and molecular signalling. Similarly, gram-positive strain is enriched with transcription factor (σ) and numerous genes which help in phage-related functions, membrane-associated biochemical reactions, glycolysis, motility and chemotactic response [28]. Various environmental conditions may downregulate or upregulate the expression of genes involved in biofilm survival and growth. Limited oxygen and nutrient supply, exposure to heavy metals and pharmacologically active molecule may downregulate gene involved in biofilm maturation and dispersal. Xu et al. investigated that RpoN and RpoN mutant genes behave in opposite directions. RpoN supports motility and acts as an antistress factor that makes hassle-free survival of *L. aggregata* under marine conditions. In contrast, the mutant gene significantly reduces biofilm formation [29]. Recent studies stressed about the environmental factors such as biotic/abiotic surface composition and surrounding microenvironment that facilitates or hinders the biofilm formation or growth via alteration in gene expressions of *SaG*, *Agr* and *MgrA* genes in *Staphylococcus aureus* [30, 31].

25.2.5 Material Surface Topography Affecting Bacterial Adhesion and Survival

Bacterial attachment on the surface depends on its topography that permits a series of biofilm formation series, i.e. initiation to dispersal. Several factors that might influence bacterial implantation are the following:

(a) *Surface Roughness*

Bacteria move through Brownian movement, gravitational and hydrodynamic forces as well as its motility activity. Bacteria adhere to material surface via various electrostatic interactions, Van der Waal forces and acid-base interactions. Bacteria adhere to the surface with the help of microscopical appendages (pili) which permit bacteria to conquer energy barrier and cling to the surface. Crest and trough sites on the surface endorse bacterial strain to retain it and stimulate colonization [32]. Recently, zirconium [33], titanium/polyethylene [34] and cobalt-chromium alloy [35] for dental, orthopaedic, and hip implants are used for prevention against microbial growth.

(b) *Surface Porosity*

Surface porosity has been declared as the significant aspect that allows bacterial retention as well as potentiates its survival. Bacteria preferentially adhere to the substrate with high porosity and grooves which provide large surface area for bacteria colonization [36]. Interestingly, the small pore size of surface in comparison to microbial cell size would not allow to the attachment of bacterial cell and initiate biofilm formation. Mechanistically, nanorange pore size generates energetic repulsive forces between bacterial cell and surface that ultimately result in impairment in biofilm formation [37]. The concept of designing nanoscale pore size substrate such as alumina with 15–20 nm range inhibiting bacterial adhesion [37], and bioinspired nanostructured surfaces, has been studied for limiting bacterial adherence as well as motility [38].

(c) *Surface Hydrophilicity/Hydrophobicity*

Surface hydrophilicity and hydrophobicity are the central facet that influences bacterial adherence. Surface wettability advocates for the extent of bacterial interaction. Ultra-hydrophobic surfaces (static contact angle $>150^\circ$) repel water droplets; thus they do not entertain bacteria for adhesion [39]. This might occur due to nanoscale roughness and shear stress that limit bacterial contact. Cell surface hydrophobicity (CSH) allows microbial granules to attach to organic droplets and thus decompose organic/hydrocarbon pollutants as well as enhance bioremediation [40]. Mainly, gram-negative bacteria release membrane vesicles which inflate membrane hydrophobicity, thus allowing their contact with the surface. Tanaka et al. reported that biofilm contributors such as EPS, phospholipids, nucleic acids and proteins impart hydrophobic or hydrophilic surface [41].

(d) *Surface Charge and Energy*

Surface geometry and zeta potential reduce the attachment of bacterial cell, but attachment depends on bacterial cell surface hydrophobicity, surface charge, presence of pili and fimbriae. The interesting point is the substrate having positive charge can harm bacterial cell membrane, especially the gram-negative one despite gram-positive, thus affecting bacterial mobilization. This outcome might be due to the structural difference between the two strains. “Charge regulation” along with electrostatic potential leads to variation in pH of the surrounding. Bacterial contact with the surface causes pH alteration leading to variation in surface charge, which advocates that net surface zeta potential might affect bacterial cellular bioenergetics [42].

25.3 Development of Diverse Biologically Synthesized Nanostructures

Biologically derived nanostructures have been used from several years as therapeutic, diagnostic, theranostic, antimicrobial and drug delivery agent and in foremost diverse biomedical applications. The main motive for biogenic derivation belongs to lesser side effects, high therapeutic window and compatibility with host cells. In the current state of the art, biological approaches have implicated different sub-methods

(using enzymes, phages, bacterial proteins and nucleic acids) for synthesizing nanostructures as well as functionalized coated biomaterials that can be used for combating microbial contamination and limiting hazardous consequences for the human population and environment. Some recent methodologies have been described in the following subsections.

25.3.1 Enzymatically Synthesized Nanoparticles

Enzymes are well known for limiting the growth of biofilm formation. Plausibly, they hinder the functionalization of bioadhesive polymer and disruption of biofilm scaffold, imparting biocidal activity as well as hampering cell-to-cell communication. Yeon et al. stated that glucose oxidase-immobilized chitosan nanoparticles (CS-NPs) advocated antibiofilm as well as antibacterial action via generating H_2O_2 (inducer of oxidative stress) through the conversion of glucose within the bacterial cell. CS-NPs have synthesized using cross-linking and enzyme precipitation methods [43]. Chen et al. reported the conversion mechanism of soluble palladium into metallic palladium nanoparticle via H_2 (enzymatic/autocatalytic effect) using H_2 -based membrane biofilm reactor. In this approach, membrane biofilm bioreactor infuses H_2 in the biofilm that leads to microbial reduction of Pd(II) to controllable Pd^0 via electron transfer, thus utilizing for wastewater treatment, removal of contaminants from water resources [44]. Nowadays, nanozymes (nanoparticles that mimic natural enzymes) have gathered an attraction for their multidimensional properties. H_2O_2 inspired iron oxide nanoparticles have been explored for disruption of biofilm via ROS production; vanadium pentoxide nanowires used for water treatment in oceans, catalytic nanoparticle integrated with H_2O_2 for dental plaque, AuNP micelle with Ce(IV) in centre adsorbed on ferrosiferic oxide/silicon dioxide shells which mimic DNase enzyme thus hinder the growth [45]. Halogenating enzymes that might diffuse into the bacterial membrane thus show cidal effect. AHL analogs have been used as major tools for combating biofilm formation such that halogenated furanones hinder AI-2 pathways, while brominated furanones disrupt AI-2 generating *LuxS* enzyme. Haloperoxidases catalyse oxidative halogenation of bacterial cell signalling compounds that hamper biofilm survival [46]. Liu et al. stated that metal-organic framework and cerium (IV) nanozyme were designed to mimic peroxidases and DNase-like action, thus helping in dissolving biofilm matrix-forming nucleic acids, eDNA and proteins [47].

25.3.2 Phage-Embedded Nanofibres and Nanoparticles

Phage-assisted nanoform synthesis has been radically explored in the field of nanoengineering. Lytic phages (which destroy the host cell machinery) are always considered as therapeutic agents for the treatment of mammalian as well as plant disease. Bacteriophages belong to the class of virus that cling on the bacterial host cell, multiply and survive. They persist in receptor-specific binding in the host cell.

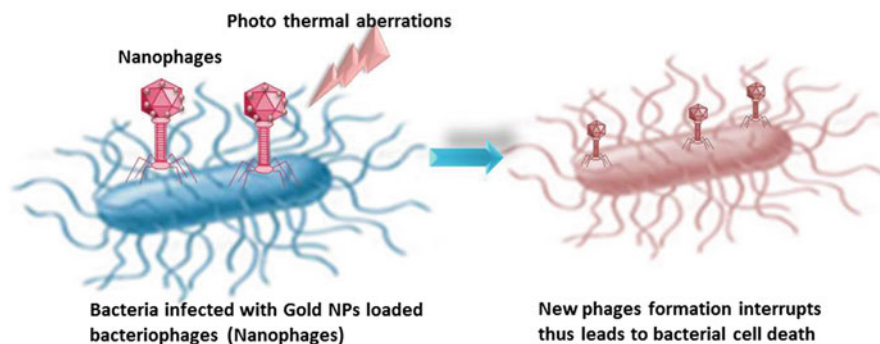


Fig. 25.1 Diagrammatic representation of bactericidal mechanism by gold nanophages

Nanoform synthesis using phage approach is the best among green strategies. Different nanostructures have been explored, such as cobalt nanowires using M13 phage, Au hybrid-Co nanowire obtained using binding protein inculcated in the filament of virus resultant nanoform persist potent antibiofilm activity [48]. Bacteriophage mechanism behind the aetiology for biofilm inhibition involves digestion of cell membrane by the lytic enzyme as well as binding with receptor-binding proteins. Peng et al. examined the antibacterial potential as well as antibiofilm potential using phage-inspired nanorods. Phage has been conjugated with gold nanorods to form phanorods. These phage-assisted nanorods are advanced form than phage therapy for combating antibiotic-resistant infections as well as biofilm interruptions [49].

Phage nanorods become an exciting tool for preventing biofilm growth. The concept behind designing phage NR is to infect bacterial host cell via inflation in the exponential growth of virions and conjugated NR especially gold NR used for imparting infrared light which cause phage destruction and hamper replication that results in bacterial cell death [50]. Diagrammatic representation of Gold nanophage mechanism described in Fig. 25.1.

25.3.3 Phytonanotechnologically Synthesized Metallic Nanostructures

Phytonanotechnology has gathered an attraction around the globe due to its numerous applications and lesser side effects along with less monetary investment and single-step strategy. Preceding articles mentioned that phytochemicals such as flavonoids, polyphenols, alkaloids, anthraquinone, amides, terpenoids, glycosides, rhamnosides and saponins serve as reducing/protecting/capping/stabilizing agents. Bioactive molecules in natural extract are used for formation of metal/metalloid/metal oxide nanoparticles [51]. Silver nanoparticles have been used from several years as antibacterial and agent, but in recent studies, their antibiofilm potential has been explored [52]. Herein, the antibacterial inhibition potential is higher in gram-negative than gram-positive strain as well as in comparison to gold standard

gentamicin sulphate. The antibacterial action addressed a dose-dependent effect on both strains. Mechanistically, AgNPs act differently against both bacterial strains that might be due to the difference in the structural composition of the bacterial cellular membrane. The difference in the thickness of the peptidoglycan layer between gram-negative and gram-positive bacteria is one of the plausibility for AgNP penetration. Gram-negative bacteria have a negative charge on cell wall surface due to the presence of carboxyl, amino and phosphate groups, thus electrostatically attracting positive ion H^+ and Ag^+ present in the surrounding environment, which directly poses toxic effects on bacterial cell via interacting with S-H group of cysteine and thus inhibiting the functioning of specific enzyme and electron transport mechanism. Small nanoparticles penetrate easily and generate ROS within the bacterial environment, interrupt subcellular reactions, damage nucleic acid, inactivate biomolecules and hinder the exchange for waste and nutrient [53]. Gram-positive bacteria defend themselves against oxidative stress using the disulfide reductase enzyme (thioredoxin system *Trx*) mechanism. Herein, AgNPs bind to active sites such as *Trx* system and block enzymatic action. This process leads to an imbalance in thioredoxin function which directly leads to bacterial death via inflation in the level of ROS [54]. Selenium nanoparticles [55, 56], gold nanoparticles [57, 58] and titanium oxide nanoparticles inhibit rhamnolipids (responsible for motility), thus hampering biofilm survival [59]. Iron oxide nanoparticles combat biofilm formation via oxidative stress (ROS) production [60]. Copper nanoparticles (ISQ/CAS@CuNPs) capped with glycosides and isoquercetin impart membrane-damaging effect and reduce cell surface hydrophobicity, thus hindering stable attachment on the surface and ultimately inhibit biofilm formation [61]. Different nanostructures act through their peculiar pathways for combating biofilm formation, as depicted in Fig. 25.2.

25.3.4 Antimicrobial Peptide (AMP) Nanosheet Scaffolds and Hydrogels

Antimicrobial peptides and proteins (AMP) immobilized on the surface or conjugated with antibiotics are host-defensive molecules that demonstrate broad spectrum of activity against bacteria, fungi and yeast. AMPs have been known for their immunomodulatory effects, promoting wound healing and their various multidimensional applications. Acosta et al. designed a unique self-assembled monolayer nanocoating including AMPs (GL13K) along with elastin-like recombinamers (ELRs). Recombinant AMP/polypeptide monolayer via covalent interactions was immobilized on gold surface, and antibiofilm efficiency was examined against *S. aureus* and *S. epidermidis*. GL13K adhere to biomaterial surface and act as front runners to disrupt the mature biofilm and minimize bacterial colonization and are thus used immensely for coating indwelling medical devices inhibiting biofilm-associated infections [62].

Interestingly, natural AMPs do not possess antibacterial property compared to antibiotics; simultaneously they persist with more side effects. For overcoming this

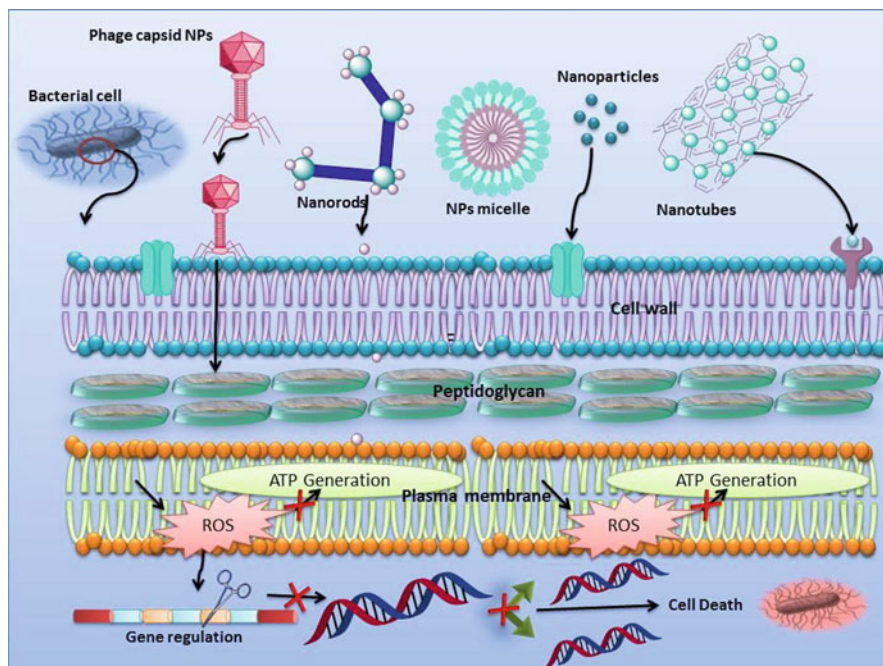


Fig. 25.2 Diagrammatic illustration of different nanostructures acting through their peculiar pathways for combating biofilm formation

situation, recombinant AMPs have been designed, and for enhancing its activity, it has been loaded/self-assembled on nanostructured hydrogel without cytotoxic effect. The plausible mechanism behind this approach might be interruptions in cell wall synthesis, inhibiting the nucleic acid synthesis and protein via interfering DNA binding and hindering enzymatic actions and activation of autolysin (self-degrading action) [63]. Khan et al. enumerated that AMP (epsilon-poly-L-Lysine) conjugated with catechol based hydrogel via cross-linking inspired by the interaction between amino and phenol group (mussel inspired chemistry) were designed and examined for antimicrobial action against multidrug resistant *Acinetobacter baumannii* for preventing infections during burn condition [64]. Rozenbaum et al. explored the efficiency of monolaurin lipid nanocapsules (MN-LNCs) adsorbed with antimicrobial peptide (AMP) against biofilm growth and for wound healing. They explored the synergistic effect of MN-LNCs@AMP against floating *Staphylococcus aureus* via membrane disruption mechanism. This mechanism was achieved by controlled penetration of antimicrobial peptide and antibacterial nanocapsules beneath the dense matrix of biofilm. MN-LNCs@AMP has been an outstanding combination owing to its wound healing capacity along with antibacterial efficacy [65].

25.3.5 Novel Antibiofilm Coating Using Polymeric Nanoparticles

Nowadays, polymeric nanoparticle coating is vastly used to develop unique antibacterial smart surfaces for overcoming the challenge of bacterial adhesion and biofilm formation. In the current scenario, numerous polymer nanoparticle-decorated surfaces have been developed such as polydopamine (PDA) (biopolymer) supramolecular assembly with poly(N,N-dimethylacrylamide) (PDMA) hydrophilic polymer to form a uniform and stable coating for biofilm inhibition. PDA-PDMA nanocoating effectively hampers biofilm formation as well as bacterial adhesion at their early stage and further colonization [66]. PDA coating grafted PEG along with the formation of NO precursor that releases nitric oxide that induces bacterial dispersal and killing, moreover averting bacterial attachment to the surfaces [67]. Epoxy/titanium oxide (TiO₂) and epoxy/Ag-doped -TiO₂ nanocomposite coatings have been exploited for biofilm disruption [68], halloysite nanotubes (HNTs) were loaded with carvacrol for sustained-release. Further, these loaded tubes grafted into polyurethane nanocomposite coating for antibiofilm efficiency via depolarising cytoplasmic membrane of pathogenic bacteria, especially gram-negative bacteria, thus demolishing the survival of mature biofilm [69]. Surface functionalization with polymeric nanoparticles/nanocomposite coating advocates for antifouling potential; hinders bacterial adhesion via altering surface geometry, membrane disruption, charge difference and enzymatic effect; and poses antibiofilm activity against mature biofilm along with bare minimum side effects. These smart surfaces would have been used in combating nosocomial infection, which is a prominent issue around the globe.

25.4 Nanostructures as Biofilm Inhibitor

Interaction between nanostructures and biofilm surfaces has been explored vastly in recent scenario among microbial nanobiotechnology research area. Nowadays, biofilm-associated infections become a significant issue as well as biofilm formation on the water resources and construction area, aggravating corrosion and fouling. For combating this situation, research has been done in the direction to alter surface geometry, hindering metabolic waste exchange, affecting the epigenomic environment of residing bacteria, etc.

25.4.1 Altering Surface Topography as an Initial Weapon

Nanoforms proved themselves a best alternative for antibiotic as bacterial population quickly develop resistance. But in the case of nanostructures, it is challenging to build resistance for bacterial cell due to their cell permeability factor. Silver, gold, copper and iron-based nanoparticles affect biofilm formation due to their large surface area-to-mass ratio. The antibacterial coating on cardiac/dental implants [70] and urinary, intravascular [71] and neurosurgical catheters/bone implants

make them resistant for the bacterial adhesion [72]. Antibacterial surfaces involve metal oxide nanocoating, nanopolymer, nanocomposite coating, and multiple NPs grafted with antibiotics. This nanocoating changes surface chemistry and physicochemical properties of the surface that prevent adhesion of bacteria and cellular disruption along with enhancing cellular adhesion, protein adhesion and bioactive agent delivery via surface permeation [73]. Production of nanopattern is the most recent approach using lithography (majorly nanoimprint lithography) that involves pattern accepting from the master surface. This technique creates a bioactive surface with modulating surface chemistry that advocates for antifouling and antibiofilm efficiency [74]. Zwitterionic surfaces contain an equal number of positive and negative charges, thus leading to electrical neutrality, which results in enhanced hydrophobicity; this hydrophobicity prevents bacterial contact and has an antifouling action [75]. “Nanoantibiotic” is used for combating resistance issue as well as shielding active drug from degradation by the enzymatic/environmental effect. Nanoantibiotics act via interrupting protein synthesis, DNA replication, gene transfer process and transcription and translation in the bacterial epigenetic atmosphere [76].

25.4.2 NPs Affecting Epigenetic Atmosphere of Bacterial Community

Epigenomic study reveals the chemical compounds that modify gene expression depicted by the bacterial community, such as DNA methylation and micro non-coding RNA-based mechanisms, lacking histone/nucleosomes and possessing gene-regulating phase variation for maintaining heterogeneity [77]. DNA adenosine methyltransferase in *E. coli* is essential for expression for a gene like *pap* operon [25]. DNA methylation has an essential role in cell development and gene regulation and is involved in virulence of bacteria. Inhibition of catalysis of methyl group transfer will lead to a reduction in pathogenicity in bacteria. Basically, identification of crucial molecules for determining pathogenicity within bacteria is the result of “omic technologies” that give rise to the construction of “nanobiosensors”. These biosensors support molecular diagnostic in virulent strains and pathological state of the disease. Recently, nanomaterials have been exploited for fabricating efficient biosensors to detect DNA methyltransferase assay [78]. Nanomaterial induces alterations in DNA methylation, histone acetylation/methylation and micro small non-coding RNA. Silver nanoparticles advocate for suppression of gene that encodes for oxidative damage repair system of bacterial DNA especially in (*mutM*, *mutS*; DNA & *nth*, *xthA*; repair enzymes) *Escherichia coli*. AgNP-treated bacterial cell upregulates the gene expressing sulphur influx, protein for magnesium and copper transport as well as antioxidant value [79].

25.5 Role of Nanoparticles in Dealing with Different Epigenomic Markers for Hindering Multistep for Biofilm Formation to Dispersal

In the current state of the art, nanostructures have been exploited for the construction of detecting molecules/sensing agent for biofilm epigenetic markers that proved as the backbone for bacterial community survival beneath dense matrix. DNA methylation is essential for maintaining biological function such as transcription, DNA replication, nascent progeny formation, DNA repair system, plasmid transfer, cell motility and bacterial cell communication. Two critical enzymes (methyltransferase and endonuclease) act as defenders for bacterial infection against phage attack [80].

Nanopore sequencing is upcoming third-generation sequencing technologies that utilize nanopore technology as well as nanostructures for detecting epigenetic markers for determining the extent of pathogenicity of biofilm-forming bacterial strains. There are several novel techniques for detecting methylation process [81]. Wu et al. described that graphene nanoribbon, nanogap, nanosheet and nanopore have the potential for detecting DNA and genomic activities, due to its high sensitivity, ultrathin thickness, large surface area (honeycomb structure), excellent compatibility and easy to functionalize with chemical and biological site. Owing to these properties, graphene is used as a biosensing nanomaterial [82]. Chen and Liu revealed that metal NPs are used to fabricate conical nanopore and utilized as biosensors for detecting alteration in epigenetic makeup for bacterial cellular integrity [83]. Hiraoka et al. mentioned that nanopore sequencing has been also used for the metaepigenomic analysis of various novel motifs present in the enzymes responsible for DNA methylation and genes involved in small non-coding RNA detection in bacteria [84].

25.6 Future Concerns for Inculcating Nanocargos as Front Runners for Treating Biofilm-Associated Pathological Conditions

Biofilm-associated infections are increasing day by day in the current situation all over the world. It became a serious issue of concern; the majority of infections are caused by indwelling devices such as contact lenses (keratitis), urinary (urinary tract infection)/central venous catheter, heart valve/pacemaker (prosthetic valve endocarditis), endotracheal tubes (pneumonia) and hip (gingivitis) and breast implant surfaces. The most prominent bacterial strains involved in biofilm-forming ability in the lumen of the catheter, the surface of contact lenses, valve/pacemaker surface, endotracheal tubes and orthopaedic and breast implant surfaces are *Staphylococcus aureus*, *Staphylococcus epidermidis*, *Candida* and *Proteus* strains, *Pseudomonas aeruginosa*, *Escherichia coli* and *Klebsiella pneumoniae* [85].

As far as nanotechnology aspect is concerned in the current state of affairs, different nanoforms have been scrutinized for combating biofilm-associated infections. There are several pathways by which nanostructures can overcome

bacterial adhesion to biotic as well as abiotic surface. As discussed in earlier sections, surface functionalization and alteration in surface chemistry are the initial steps for inhibiting bacterial contact. NPs/nanoclusters with the inherent antibacterial property such as silver, gold and metal oxide NPs (i.e. magnesium oxide, iron oxide, copper oxide, zinc oxide, titanium dioxide) act through ROS formation, upregulation of oxidative stress markers, inhibition of respiratory chain (electron transport chain) and ROS generation to obstruct energy formation by impeding ATP generation [75].

Recent advancement in the field of nanobiotechnology is the discovery of “nanozymes” which are generally known as artificial enzymes, nanoforms that contain intrinsic enzyme-like efficiency. Nanozymes offer numerous benefits over naturally occurring enzymes. These benefits comprise easy synthesis, cost-effectiveness, stability, and vigorous catalytic efficiency, simple methods of synthesis, low cost, high stability, catalytic performance and nanostructure surface functionalization. In reference of these advantages, nanozymes are enormously being investigated for determining an array of applications such as biosensing ability, agents in immunoassays, disease diagnostics and therapeutic agents, theranostic agents, bioremediator and prevention from oxidative stress [86]. Khulbe et al. reported that ceria-based nanoenzyme and polyacrylic acid decorated ceria nanoenzymes, active against biofilm formation via initially targeting bacterial cell membrane and further hydrolysing phospholipid layer present in the cellular membrane [87].

Nanocapsules are the weapon for novel drug delivery system in which lipophilic and hydrophilic drugs were coencapsulated and shielded with the polymeric membrane, which helps in the penetration of active agent in the deeper target area [88]. Rozenbaum et al. explored the efficiency of monolaurin lipid nanocapsules (MN-LNCs) adsorbed with antimicrobial peptide (AMP) against biofilm growth and for wound healing. They investigated the synergistic effect of MN-LNCs@AMP against floating *Staphylococcus aureus* via membrane disruption mechanism. This mechanism was achieved by controlled penetration of antimicrobial peptide and antibacterial nanocapsules beneath the dense matrix of biofilm. MN-LNCs@AMP has been an outstanding combination owing to its wound healing capacity along with antibacterial efficacy [65]. Dalcin et al. stated that dihydromyricetin (flavonoid)-loaded nanocapsules were used to overcome the biofilm formation of *Pseudomonas aeruginosa* (PA) in the urinary catheter. They reported excellent antibiofilm activity against *P. aeruginosa* [89].

Nanoemulsion has been recently used for preventing biofilm formation using natural essential oils having antibacterial activity. Prateeksha et al. reported that essential oil containing eugenol and methyl salicylate-based nanoemulsion showed antibiofilm activity and reduction in virulence via downregulation of cellular signalling gene expression, which encodes for *LuxR* and *LuxS* [90]. Ramalingam and Lee reported the antibiofilm potential of EDTA-containing nanoemulsion (coated on glass surface) prevents against multidrug-resistant (MDR) *Acinetobacter baumannii*. The mechanism behind the activity includes a reduction in surface lipopolysaccharide (LPS) and bacterial integration [85].

Future perspective of nanobiotechnology for biofilm-associated pathological conditions might involve the discovery of phage-encoded nanocomposite, nanoemulsion for the alteration of surface topography, nanorod developments, nanoparticles as carrier for antimicrobial compounds and the formation of antimicrobial peptide scaffolds on the surface for preventing bacterial contact by altering surface roughness. These approaches help in lowering adverse effects and maintaining proper therapeutic window along with decreasing bacterial tolerance. Nanoapproaches are inculcated in the current era due to their infinite advantages as mentioned earlier; one of the main benefits is its non-immunogenicity and maximized pharmacological effects.

25.7 Conclusion and Outlook

In the current state of the art, biofilm formation issue is the most burgeoning area of concern around the globe. Somehow, biofilm is suitable for some surfaces regarding prevention from the harsh environment, but herein, we have discussed that bacterial biofilm acts as a zombie for human as well as aquatic, terrestrial ecosystem. Bacteria adhere to biotic/abiotic surface, especially with high hydrophobicity, roughness and highly porous property. Bacterial adhesion follows a series of steps for biofilm formation, starting from the adhesion to matrix formation (EPS) to maturation to dispersion, and further finds a suitable place with high nutrient content required for matrix formation. At the genetic level, AHL, AI-2, AI-1, *LuxS*, *LuxR* and eDNA-regulating gene help for inhibiting biofilm growth. Downregulation of gene expressing these functions leads to biofilm inhibition and reduction in virulence of bacterial strains. Dispersal enzymes, ROS production, cellular disruption via difference in osmolarity and σ factor affect bacterial survival at the molecular level.

Nowadays, biofilm-associated infections have become a significant threat among the human population. Indwelling medical devices (urinary catheter/intravenous catheter, heart valve, contact lenses, hip implants, dental implants) are prone to get contaminated with a bacterial biofilm that further harms exposed patient. To overcome these serious issues, nanobiotechnology has been inculcated as nanoparticles, nanocapsules, nanoenzymes, nanophages, AMP nanocoating and nanoemulsion for treating surfaces for preventing bacterial adhesion. Mechanistically, these nanoforms act through altering surface geometry, cellular signalling and epigenetic makeup. This whole chapter provides informative content related to biofilm and associated mechanism along with biofilm infections overcoming strategies using the nanobiotechnological approach in a single platform.

Acknowledgement IS would like to acknowledge the Department of Science and Technology, Government of India for providing financial assistance.

References

1. Jamal M, Ahmad W, Andleeb S, Jalil F, Imran M, Nawaz MA, Hussain T, Ali M, Rafiq M, Kamil MA (2018) Bacterial biofilm and associated infections. *J Chin Med Assoc* 81(1):7–11
2. Toyofuku M, Inaba T, Kiyokawa T, Obana N, Yawata Y, Nomura N (2016) Environmental factors that shape biofilm formation. *Biosci Biotechnol Biochem* 80(1):7–12
3. Luther MK, Parente DM, Caffrey AR, Daffinee KE, Lopes VV, Martin ET, LaPlante KL (2018) Clinical and genetic risk factors for biofilm-forming *Staphylococcus aureus*. *Antimicrob Agents Chemother* 62(5):e02252–e02217
4. Zamani H, Salehzadeh A (2018) Biofilm formation in uropathogenic *Escherichia coli*: association with adhesion factor genes. *Turk J Med Sci* 48(1):162–167
5. Abad ED, Khameneh A, Vahedi L (2019) Identification phenotypic and genotypic characterization of biofilm formation in *Escherichia coli* isolated from urinary tract infections and their antibiotics resistance. *BMC Res Notes* 12(1):796
6. Karami P, Khaledi A, Mashoof RY, Yaghoobi MH, Karami M, Dastan D, Alikhani MY (2020) The correlation between biofilm formation capability and antibiotic resistance pattern in *Pseudomonas aeruginosa*. *Gene Reports* 18:100561
7. Kot B, Sytykiewicz H, Sprawka I (2018) Expression of the biofilm-associated genes in methicillin-resistant *Staphylococcus aureus* in biofilm and planktonic conditions. *Int J Mol Sci* 19(11):3487
8. Nahar S, Mizan MFR, Ha A J-w, Ha S-D (2018) Advances and future prospects of enzyme-based biofilm prevention approaches in the food industry. *Compr Rev Food Sci Food Saf* 17(6):1484–1502
9. Şahin R (2019) Extracellular enzymes, pathogenity and biofilm forming in staphylococci. *Clin Dia and Cure Open A Open J I*(1):12–17
10. Tielen P, Rosenau F, Wilhelm S, Jaeger K-E, Flemming H-C, Wingender J (2010) Extracellular enzymes affect biofilm formation of mucoid *Pseudomonas aeruginosa*. *Microbiology* 156(7):2239–2252
11. Flemming H-C, Wingender J, Szewzyk U, Steinberg P, Rice SA, Kjelleberg S (2016) Biofilms: an emergent form of bacterial life. *Nat Rev Microbiol* 14(9):563–575
12. Anderson JK, Huang JY, Wreden C, Sweeney EG, Goers J, Remington SJ, Guillemin K (2015) Chemorepulsion from the quorum signal autoinducer-2 promotes helicobacter pylori biofilm dispersal. *MBio* 6(4):e00379–e00315
13. Koo H, Allan RN, Howlin RP, Stoodley P, Hall-Stoodley L (2017) Targeting microbial biofilms: current and prospective therapeutic strategies. *Nat Rev Microbiol* 15(12):740–755
14. Hossain MA, Lee S-J, Park N-H, Mechesso AF, Birhanu BT, Kang J, Reza MA, Suh J-W, Park S-C (2017) Impact of phenolic compounds in the acyl homoserine lactone-mediated quorum sensing regulatory pathways. *Sci Rep* 7(1):10618
15. Stubbenieck RM, Straight PD (2020) Specialized metabolites for bacterial communication. In: Liu H-W, Begley TP (eds) *Comprehensive natural products III*, vol 3. Elsevier, Oxford, pp 66–96
16. Yi J, Zhang D, Cheng Y, Tan J, Luo Y (2019) The impact of *Paenibacillus polymyxa* HY96-2 luxS on biofilm formation and control of tomato bacterial wilt. *Appl Microbiol Biotechnol* 103(23–24):9643–9657
17. Fleming D, Rumbaugh KP (2017) Approaches to dispersing medical biofilms. *Microorganisms* 5(2):15
18. Fleming D, Chahin L, Rumbaugh K (2017) Glycoside hydrolases degrade polymicrobial bacterial biofilms in wounds. *Antimicrob Agents Chemother* 61(2):e01998–e01916
19. Sharma K, Singh AP (2018) Antibiofilm effect of DNase against single and mixed species biofilm. *Foods* 7(3):42
20. Kaplan JB, Mlynek KD, Hettiarachchi H, Alamneh YA, Biggemann L, Zurawski DV, Black CC, Bane CE, Kim RK, Granick MS (2018) Extracellular polymeric substance (EPS)-degrading

- enzymes reduce staphylococcal surface attachment and biocide resistance on pig skin in vivo. *PLoS One* 13(10):e0205526
21. Wolska KI, Grudniak AM, Rudnicka Z, Markowska K (2016) Genetic control of bacterial biofilms. *J Appl Genet* 57(2):225–238
 22. Taylor PK, Van Kessel AT, Colavita A, Hancock RE, Mah T-F (2017) A novel small RNA is important for biofilm formation and pathogenicity in *Pseudomonas aeruginosa*. *PLoS One* 12(8):e0182582
 23. Diallo I, Provost P (2020) RNA-sequencing analyses of small bacterial RNAs and their emergence as virulence factors in host-pathogen interactions. *Int J Mol Sci* 21(5):1627
 24. Paget MS (2015) Bacterial sigma factors and anti-sigma factors: structure, function and distribution. *Biomol Ther* 5(3):1245–1265
 25. Adhikari S, Curtis PD (2016) DNA methyltransferases and epigenetic regulation in bacteria. *FEMS Microbiol Rev* 40(5):575–591
 26. Saha A, Arora K, Sajid A, Arora G (2018) Cellular signaling in bacterial biofilms. In: Implication of quorum sensing system in biofilm formation and virulence. Springer, Cham, pp 81–109
 27. Suryawati B (2018) Zinc homeostasis mechanism and its role in bacterial virulence capacity. *AIP Conf Proc* 1:070021–070027
 28. Burianni G, Donati I, Cellini A, Fiorentini L, Vanneste J, Spinelli F (2017) Molecular signalling in *Pseudomonas syringae* pv. actinidiae. IX International Symposium on Kiwifruit 1218:299–306
 29. Xu T, Yu M, Liu J, Lin H, Liang J, Zhang X-H (2019) Role of RpoN from *Labrenzia aggregata* LZB033 (Rhodobacteraceae) in formation of flagella and biofilms, motility, and environmental adaptation. *Appl Environ Microbiol* 85(7):e02844–e02818
 30. Kim BR, Bae YM, Lee SY (2016) Effect of environmental conditions on biofilm formation and related characteristics of *Staphylococcus aureus*. *J Food Saf* 36(3):412–422
 31. Schilcher K, Horswill AR (2020) Staphylococcal biofilm development: structure, regulation, and treatment strategies. *Microbiol Mol Biol Rev* 84(3):e00026–e00019
 32. Ammar Y, Swailes D, Bridgens B, Chen J (2015) Influence of surface roughness on the initial formation of biofilm. *Surf Coat Technol* 284:410–416
 33. Yu P, Wang C, Zhou J, Jiang L, Xue J, Li W (2016) Influence of surface properties on adhesion forces and attachment of *Streptococcus mutans* to zirconia in vitro. *Biomed Res Int* 2016:8901253
 34. Malhotra R, Dhawan B, Garg B, Shankar V, Nag TC (2019) A comparison of bacterial adhesion and biofilm formation on commonly used orthopaedic metal implant materials: an in vitro study. *Indian J Orthop* 53(1):148–153
 35. Aherwar A, Singh AK, Patnaik A (2016) Cobalt based alloy: a better choice biomaterial for hip implants. *Trends Biomater Artif Organs* 30(1):50–55
 36. Khelissa SO, Abdallah M, Jama C, Faille C, Chihib N-E (2017) Bacterial contamination and biofilm formation on abiotic surfaces and strategies to overcome their persistence. *J Mater Environ Sci* 8(9):3326–3346
 37. Feng G, Cheng Y, Wang S-Y, Borca-Tasciuc DA, Worobo RW, Moraru CI (2015) Bacterial attachment and biofilm formation on surfaces are reduced by small-diameter nanoscale pores: how small is small enough? *NPJ Biofilms Microbiomes* 1(1):15022
 38. Tripathy A, Sen P, Su B, Briscoe WH (2017) Natural and bioinspired nanostructured bactericidal surfaces. *Adv Colloid Interf Sci* 248:85–104
 39. Achinas S, Charalampogiannis N, Euverink GJW (2019) A brief recap of microbial adhesion and biofilms. *Appl Sci* 9(14):2801
 40. Talukdar P, Sharma C, Doley A, Baruah K, Borah A, Agarwal P, Deori P (2017) Isolation and characterization of biosurfactant producing microorganisms from petroleum contaminated soil samples for EOR and bioremediation. *Petrol Sci Tech* 35(22):2102–2108
 41. Tanaka N, Kogo T, Hirai N, Ogawa A, Kanematsu H, Takahara J, Awazu A, Fujita N, Haruzono Y, Ichida S (2019) In-situ detection based on the biofilm hydrophilicity for environmental biofilm formation. *Sci Rep* 9(1):8070

42. Zhu H (2016) Impact of the charge-regulation effect on bacterial growth and bioavailability of sorbed growth substrates. Theses and dissertations, 2913
43. Yeon K-M, You J, Adhikari MD, Hong S-G, Lee I, Kim HS, Kim LN, Nam J, Kwon S-J, Kim MI (2019) Enzyme-immobilized chitosan nanoparticles as environmentally friendly and highly effective antimicrobial agents. *Biomacromolecules* 20(7):2477–2485
44. Zhou C, Ontiveros-Valencia A, Wang Z, Maldonado J, Zhao H-P, Krajmalnik-Brown R, Rittmann BE (2016) Palladium recovery in a H₂-based membrane biofilm reactor: formation of Pd(0) nanoparticles through enzymatic and autocatalytic reductions. *Environ Sci Technol* 50(5):2546–2555
45. Cormode DP, Gao L, Koo H (2018) Emerging biomedical applications of enzyme-like catalytic nanomaterials. *Trends Biotechnol* 36(1):15–29
46. Herget K, Frerichs H, Pfitzner F, Tahir MN, Tremel W (2020) Functional enzyme mimics for oxidative halogenation reactions that combat biofilm formation. In: *Nanozymology*. Springer, Singapore, pp 195–278
47. Liu Z, Wang F, Ren J, Qu X (2019) A series of MOF/Ce-based nanozymes with dual enzyme-like activity disrupting biofilms and hindering recolonization of bacteria. *Biomaterials* 208:21–31
48. Ahiwale S, Bankar A, Tagunde S, Kapadnis B (2017) A bacteriophage mediated gold nanoparticles synthesis and their anti-biofilm activity. *Indian J Microbiol* 57(2):188–194
49. Peng H, Borg RE, Dow LP, Pruitt BL, Chen IA (2020) Controlled phage therapy by photothermal ablation of specific bacterial species using gold nanorods targeted by chimeric phages. *Proc Natl Acad Sci* 117(4):1951–1961
50. Ferriol-González C, Domingo-Calap P (2020) Phages for biofilm removal. *Antibiotics* 9(5):268
51. Singh J, Dutta T, Kim K-H, Rawat M, Samddar P, Kumar P (2018) Green synthesis of metals and their oxide nanoparticles: applications for environmental remediation. *J Nanobiotechnol* 16(1):84
52. Jinu U, Jayalakshmi N, Anbu AS, Mahendran D, Sahi S, Venkatachalam P (2017) Biofabrication of cubic phase silver nanoparticles loaded with phytochemicals from *Solanum nigrum* leaf extracts for potential antibacterial, antibiofilm and antioxidant activities against MDR human pathogens. *J Clust Sci* 28(1):489–505
53. Slavin YN, Asnis J, Häfeli UO, Bach H (2017) Metal nanoparticles: understanding the mechanisms behind antibacterial activity. *J Nanobiotechnol* 15(1):65
54. Qing Y, Cheng L, Li R, Liu G, Zhang Y, Tang X, Wang J, Liu H, Qin Y (2018) Potential antibacterial mechanism of silver nanoparticles and the optimization of orthopedic implants by advanced modification technologies. *Int J Nanomedicine* 13:3311–3327
55. Khiralla GM, El-Deeb BA (2015) Antimicrobial and antibiofilm effects of selenium nanoparticles on some foodborne pathogens. *LWT-Food Sci Technol* 63(2):1001–1007
56. Fardsadegh B, Jafarizadeh-Malmiri H (2019) Aloe vera leaf extract mediated green synthesis of selenium nanoparticles and assessment of their in vitro antimicrobial activity against spoilage fungi and pathogenic bacteria strains. *Green Process Synth* 8(1):399–407
57. Singh P, Pandit S, Beshay M, Mokkapati V, Garnaes J, Olsson ME, Sultan A, Mackevica A, Mateiu RV, Lütken H (2018) Anti-biofilm effects of gold and silver nanoparticles synthesized by the *Rhodiola rosea* rhizome extracts. *Artif Cells Nanomed Biotechnol* 46(sup3):S886–S899
58. Aljabali AA, Akkam Y, Al Zoubi MS, Al-Batayneh KM, Al-Trad B, Abo Alrob O, Alkilany AM, Benamara M, Evans DJ (2018) Synthesis of gold nanoparticles using leaf extract of *Ziziphus zizyphus* and their antimicrobial activity. *Nano* 8(3):174
59. Rajkumari J, Magdalane CM, Siddhardha B, Madhavan J, Ramalingam G, Al-Dhabi NA, Arasu MV, Ghilan A, Duraipandiayan V, Kaviyarasu K (2019) Synthesis of titanium oxide nanoparticles using Aloe barbadensis mill and evaluation of its antibiofilm potential against *Pseudomonas aeruginosa* PAO1. *J Photochem Photobiol B* 201:111667
60. Erci F, Cakir-Koc R (2020) Rapid green synthesis of noncytotoxic iron oxide nanoparticles using aqueous leaf extract of *Thymbra spicata* and evaluation of their antibacterial, antibiofilm, and antioxidant activity. *Inorg Nano-Met Chem* 51(5):683–692

61. Lotha R, Shamprasad BR, Sundaramoorthy NS, Nagarajan S, Sivasubramanian A (2019) Biogenic phytochemicals (cassinopin and isoquercetin) capped copper nanoparticles (ISQ/CAS@ CuNPs) inhibits MRSA biofilms. *Microb Pathog* 132:178–187
62. Acosta S, Quintanilla L, Alonso M, Aparicio C, Rodríguez-Cabello JC (2019) Recombinant AMP/polypeptide self-assembled monolayers with synergistic antimicrobial properties for bacterial strains of medical relevance. *ACS Biomater-Sci Eng* 5(9):4708–4716
63. Yang K, Han Q, Chen B, Zheng Y, Zhang K, Li Q, Wang J (2018) Antimicrobial hydrogels: promising materials for medical application. *Int J Nanomedicine* 13:2217–2263
64. Khan A, Xu M, Wang T, You C, Wang X, Ren H, Zhou H, Khan A, Han C, Li P (2019) Catechol cross-linked antimicrobial peptide hydrogels prevent multidrug-resistant *Acinetobacter baumannii* infection in burn wounds. *Biosci Rep* 39(6):BSR20190504
65. Rozenbaum RT, Su L, Umerska A, Eveillard M, Håkansson J, Mahlapuu M, Huang F, Liu J, Zhang Z, Shi L (2019) Antimicrobial synergy of monolaurin lipid nanocapsules with adsorbed antimicrobial peptides against *Staphylococcus aureus* biofilms in vitro is absent in vivo. *J Control Release* 293:73–83
66. Mei Y, Yu K, Lo JC, Takeuchi LE, Hadesfandiari N, Yazdani-Ahmadabadi H, Brooks DE, Lange D, Kizhakkedathu JN (2018) Polymer–nanoparticle interaction as a design principle in the development of a durable ultrathin universal binary antibiofilm coating with long-term activity. *ACS Nano* 12(12):11881–11891
67. Sadrearhami Z, Shafiee FN, Ho KK, Kumar N, Krasowska M, Blencowe A, Wong EH, Boyer C (2019) Antibiofilm nitric oxide-releasing polydopamine coatings. *ACS Appl Mater Interfaces* 11(7):7320–7329
68. Santhosh MS, Natarajan K (2015) Antibiofilm activity of epoxy/Ag-TiO₂ polymer nanocomposite coatings against *Staphylococcus aureus* and *Escherichia coli*. *CoatingsTech* 5(2):95–114
69. Hendessi S, Sevinis EB, Unal S, Cebeci FC, Menciloglu YZ, Unal H (2016) Antibacterial sustained-release coatings from halloysite nanotubes/waterborne polyurethanes. *Prog Org Coat* 101:253–261
70. Rasouli R, Barhoum A, Uludag H (2018) A review of nanostructured surfaces and materials for dental implants: surface coating, patterning and functionalization for improved performance. *Biomater Sci* 6(6):1312–1338
71. Liu H, Shukla S, Vera-González N, Tharmalingam N, Mylonakis E, Fuchs BB, Shukla A (2019) Auranofin releasing antibacterial and antibiofilm polyurethane intravascular catheter coatings. *Front Cell Infect Microbiol* 9:37
72. Wang L, Hu C, Shao L (2017) The antimicrobial activity of nanoparticles: present situation and prospects for the future. *Int J Nanomedicine* 12:1227–1249
73. Narayana S, Srihari S (2019) A review on surface modifications and coatings on implants to prevent biofilm. *Regen Eng Transl Med* 6:330–346
74. Majhi S, Mishra A (2020) Modulating surface energy and surface roughness for inhibiting microbial growth. In: *Engineered antimicrobial surfaces. Materials horizons: from nature to nanomaterials*. Springer, Singapore, pp 109–121
75. Vallet-Regí M, González B, Izquierdo-Barba I (2019) Nanomaterials as promising alternative in the infection treatment. *Int J Mol Sci* 20(15):3806
76. Natan M, Banin E (2017) From Nano to micro: using nanotechnology to combat microorganisms and their multidrug resistance. *FEMS Microbiol Rev* 41(3):302–322
77. Beaulaurier J, Schadt EE, Fang G (2019) Deciphering bacterial epigenomes using modern sequencing technologies. *Nat Rev Genet* 20(3):157–172
78. Ma F, Zhang Q, C-y Z (2020) Nanomaterial-based biosensors for DNA methyltransferase assay. *J Mater Chem B* 8(16):3488–3501
79. Roy A, Bulut O, Some S, Mandal AK, Yilmaz MD (2019) Green synthesis of silver nanoparticles: biomolecule-nanoparticle organizations targeting antimicrobial activity. *RSC Adv* 9(5):2673–2702

80. Mohapatra SS, Biondi EG (2017) DNA methylation in prokaryotes: regulation and function. In: Krell T (ed) Cellular ecophysiology of microbe. Handbook of hydrocarbon and lipid microbiology. Springer, Cham. https://doi.org/10.1007/978-3-319-20796-4_23-1
81. Gouil Q, Keniry A (2019) Latest techniques to study DNA methylation. *Essays Biochem* 63 (6):639–648
82. Wu X, Mu F, Wang Y, Zhao H (2018) Graphene and graphene-based nanomaterials for DNA detection: a review. *Molecules* 23(8):2050
83. Chen Q, Liu Z (2019) Fabrication and applications of solid-state nanopores. *Sensors* 19(8):1886
84. Hiraoka S, Okazaki Y, Anda M, Toyoda A, S-i N, Iwasaki W (2019) Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an environmental prokaryotic community. *Nat Commun* 10(1):159
85. Ramasamy M, Lee J (2016) Recent nanotechnology approaches for prevention and treatment of biofilm-associated infections on medical devices. *Biomed Res Int* 2016:1851242
86. Singh S (2019) Nanomaterials exhibiting enzyme-like properties (Nanozymes): current advances and future perspectives. *Front Chem* 7:46
87. Khulbe K, Karmakar K, Ghosh S, Chandra K, Chakravorty D, Mugesh G (2020) Nanoceria-based phospholipase-mimetic cell membrane disruptive anti-biofilm agents. *ACS Appl Bio Mater* 3(7):4316–4328
88. Pathak C, Vaidya FU, Pandey SM (2019) Mechanism for development of nanobased drug delivery system. In: *Applications of targeted nano drugs and delivery systems*. Elsevier, Amsterdam, pp 35–67
89. Dalcin A, Santos C, Gündel S, Roggia I, Raffin R, Ourique A, Santos R, Gomes P (2017) Anti biofilm effect of dihydromyricetin-loaded nanocapsules on urinary catheter infected by *Pseudomonas aeruginosa*. *Colloids Surf B: Biointerfaces* 156:282–291
90. Prateeksha, Barik SK, Singh BN (2019) Nanoemulsion-loaded hydrogel coatings for inhibition of bacterial virulence and biofilm formation on solid surfaces. *Sci Rep* 9(1):6520



Malaria in the Era of Omics: Challenges and Way Forward

26

Manish Tripathi, Amit Khatri, Vaishali Lakra, Jaanvi Kaushik, and Sumit Rathore

Abstract

Malaria, a deadly disease caused by pathogen *Plasmodium*, is a global problem. After discovery of *Plasmodium* as a causative agent of malaria, the understanding of the malaria biology using conventional techniques was at slower pace as compared to the variation in the parasite, e.g., drug resistance. In the era of omics technologies, e.g., proteomics, genomics, and metabolomics, newer strategies to combat this infection have evolved. Also, these different technologies have helped in better understanding of *Plasmodium* parasite biology in detail. Omics tools involving high-throughput technologies, automation, and data mining have helped in better understanding of the pathways and key proteins required in the life cycle of the parasite and hence pathogenesis of this disease in faster, reliable, and affordable mode. In the recent times, advance in the field of omics displayed high potential in accelerating malaria research to fight the diseases. This chapter will highlight the role of omics tools in deciphering mysteries of malaria parasite biology and their applications in diagnosis, treatment, and eradication of the disease.

Keywords

Malaria · Plasmodium · Proteomics · Genomics · Drug targets · Vaccine candidate

M. Tripathi · A. Khatri · V. Lakra · J. Kaushik · S. Rathore (✉)
Department of Biotechnology, All India Institute of Medical Sciences, New Delhi, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*,
https://doi.org/10.1007/978-981-16-0691-5_26

483

26.1 Introduction

Malaria is a life-threatening disease caused by the *Plasmodium* parasite. It spreads through bite of an infected female *Anopheles* mosquito. After the identification of *Plasmodium falciparum* as the causative agent of malaria, tremendous work has been done in the field to understand the biology of the parasite. However, the conventional techniques were not fully successful in understanding the complexity of the parasite in detail. The recent revolution in the field of omics has changed the understanding of the biology of the different organism in details and has paved a path for better strategies to counter these pathogens. “Omics era” is based on a data-intensive approach to biology that relies on high-throughput techniques, data mining methodologies, automation, and tools. Advances in the field of omics have not only made it affordable but also less time-consuming as compared to its earlier version which was expensive and time-consuming.

Omics approach to understanding malaria biology is on increase in the past few years, and it shows high potential in accelerating malaria research to fight the diseases. In this chapter, we will focus on different tools of genomics, proteomics, metabolomics, and other omics tools and strategies employed in malaria research.

26.2 Genomics in the Field of Malaria Research

Genomics employs a combination of DNA sequencing, assembly, and annotation of the genome from an organism so that it can be documented for future references [1]. Advancement in the field of sequencing technology and analysis tools has helped in understanding organisms from a better perspective. The ability to identify and characterize the genetic blueprint of an organism is a critical element in our effort to target pathogens and the development of new therapeutic strategies.

Malaria parasite genome has been sequenced in 2002 under an ambitious project which lasted for 7 years and cost around \$20 million, including three international research centers to sequence the complete *P. falciparum* genome [2, 3]. Later with the advancement in the field of next-generation sequencing (NGS) and genomics, other *Plasmodium* species, which are not infecting humans, but are good disease models, have also been sequenced and are now available in the databases [4–10]. Further tools of comparative genomics helped in establishing evolutionary relationships among different species of the parasite based on their host origins [11–13]. Genomics not only helped in deciphering the genomes of these parasite species but also highlighted the variations in the field isolates of *P. falciparum* and *P. vivax* in detail [14–18].

Plasmodium genome is haploid throughout their life cycles except for a short period of [14] diploidy after fertilization in the midgut of the mosquitoes. The genome of different *Plasmodium* species is only 2–3 times bigger than *S. cerevisiae*, i.e., 25–35 Mb, comprising 14 chromosomes. *Plasmodium* also harbors a circular apicoplast genome along with several copies of the 6 kb mitochondrial genome. *Plasmodium* genome is highly AT-rich, which ranges up to 80%

among different species of the parasite. AT content of the genome is found to be higher in introns and intergenic regions than in protein-coding regions [3]. AT richness in the genome of *Plasmodium* highlights low complexity regions, simple sequence repeats, and highly skewed codon usage bias in the genome [19–21].

Comparative genomics, an important aspect of genomics, is very helpful in the understanding of the genome evolution, regulatory elements, and presence of species-specific genes in *Plasmodium*. Comparative genomics analyses using different *Plasmodium* species have revealed that genes located in the core of the chromosomes are conserved, as compared to the subtelomeric regions, which are highly species-specific [22, 23]. The majority of the genes in these telomeric regions were found to be associated with the invasion process and was found to be absent from *P. yoelii* but present in *P. falciparum* [23]. The presence of these genes in subtelomeric regions points toward the probable role of these proteins in triggering a human immune response and perhaps leading to diversifying selection due to immune pressure [23–25]. Further, these comparative genomics studies highlighted that several aspects of parasite biology are promoting genetic diversity, including secondary loss of DNA repair proteins which allows more error-prone replication and better adaption in the parasite as compared to other organisms [20].

Genomics helped in understanding details about some of the key proteins of the *Plasmodium* parasite, e.g., var genes. *Plasmodium* genome carries at least 5–150 copies of the var genes which encode PfEMP1 and are responsible for the majority of antigenic variation in the parasite [26–28]. Recombination rate of var gene is higher than an average gene in the parasite [29, 30], and genomics studies helped in revealing the extent of antigenic variation in the family, which can further help in the identification of major patterns which have been represented, and the same can be targeted. A very good example regarding the regulation of var genes came from genome sequencing studies which highlighted the presence of five different subgroups of the var gene families, i.e., UpsA to UpsE, based on the upstream flanking regions. These subgroups have been co-regulated and differ in their functions [31, 32]. Genomics studies revealed that PfEMP1 encoded by UpsA has fewer number of a cysteine residue in the DBL region [33, 34] and found to be associated with a severe form of malaria [35–37]. Further classifying var genes based on sequence identification allowed researchers to co-relate var gene associated with the different subgroups of malaria patients, e.g., pregnancy malaria is known to be associated with Var2Csa which is believed to be the most important ligand interacting to Csa, i.e., chondroitin sulfate A in the placental endothelial lining [37–39], which will be very helpful in designing new treatment strategies.

Overall genomics has been useful in malaria research since 2002, deciphering genomes of many *P. falciparum* strains and isolates identifying genes and genetic traits [40–42]. Genomics not only have helped in identifying genes associated with virulence and severity but also enabled surveillance [43, 44] and epidemiological modeling of parasite population dynamics [45]. A further recent advance which is single-cell genome sequencing [46, 47] enabled investigation of transmission dynamics and genetic diversity [48].

In short genomics studies enabled us to know a remarkable genetic diversity of parasite, which will help in the future for targeted strategies for therapy against malaria and better surveillance for drug resistance. Also, population genomics studies further have led to an efficient approach for the identification of an association between various parasite characteristics and specific genes.

26.3 Transcriptomics in Malaria Research

The transcriptome of an organism can tell us about the full range of RNA expressed by an organism. Transcriptomics has been used to identify patterns of gene expression by analyzing changes in the mRNA levels between different experimental conditions. To carry out large-scale studies to understand RNA changes, microarray technique has been employed widely, which has made it feasible to collect large gene expression datasets quickly and for reproducibility [49].

To understand the transcriptome of the malaria parasite, initially, studies started with DNA microarray [50–52] which are now being supported by recent advances in the field of RNA sequencing [53–55]. Microarray data revealed possible roles of hypothetical proteins during the life cycle of the parasite [51, 56] and an unusual cascade of transcriptional regulation [50]. This information about hypothetical genes was very welcoming because there are many unassigned genes in the parasite and regulation of various genes is poorly understood [20, 25, 56, 57].

Data from microarray studies have shown upregulation of genes associated with various pathways, i.e., immune evasion. Also, these studies allowed identification of different key interaction during host-pathogen interaction in malaria biology [58, 59] which can be utilized as potential targets for therapeutics. Microarray studies also helped in understanding the regulation of var genes, which are one of the most important factors in immune evasion strategies by the parasite. Transcriptomic analysis showed that of all only one var gene is expressed at any given time and the rest are suppressed through epigenetic gene silencing [31, 60–62]. All of these studies point toward a highly organized and structured cascade of gene expression during the asexual blood-stage cycles.

In recent reports from single-cell transcriptomic studies, it has been shown that *P. falciparum* parasites from infected individuals do not show such pattern, but have abrupt changes in the expression profile of various genes [63]. It has been shown that a group of genes is simultaneously switched on and off during asexual blood-stage development [63]. Advancement in the transcriptomic tools such as single-cell transcriptomics helped in the identification of distinct gene expression in single cells, which have been masked earlier in population-based studies and highlighted in earlier elusive development checkpoints.

Single-cell transcriptomics have helped in the identification of genes that are not dependent on programmed developmental controls but have been influenced by other factors. Effect of several factors can influence the expression in response, as has been shown in case of drugs, where nearly 59% genes displayed threefold increase in the expression levels [64]. However, previous studies were only able to

show minimal changes in the expression levels of the same genes. Further, antigenic variation [65] and host factors [52] have also been shown to influence the transcriptional profile of the parasite with the help of single-cell transcriptomics.

Transcriptomic studies in *P. falciparum* not only helped in the identification of new genes [63] but also helped in the identification of non-protein-coding RNA and their influence on the regulation of expression [66]. Non-protein-coding (npc) RNA including small structural npc-RNA, long npc-RNA, circular RNA, and npc-RNA of unknown function has been recently characterized in the parasite [67]. Identification of these different types of RNA helped in understanding the regulation of various genes in the parasite. Role of npc-RNAs in the regulation of sexual commitment [68], antigenic variation [69], chromatin assembly, and novel parasite-specific proteases process [66] has been shown.

To date, no microRNA has been identified in the parasite [70, 71] which can be attributed to the absence of RNA interference machinery components in the parasite [72]. Genomics and transcriptomics in combination have recently been used to identify leucocyte-associated immunoglobulin-like receptor 1 (LAIR1) as an important molecule for determining susceptibility to severe malarial anemia (SMA), which is the leading cause of death in children [73].

Sexual commitment in parasite is an important step of the life cycle. Single-cell RNA sequencing reveals signature of sexual commitment in malaria parasite in which it highlighted the role of AP2-G transcriptional factor and its role in specifically upregulating regulators of gene expression and nucleosome positioning [74, 75]. Single transcriptomic analysis of individual parasite allowed precisely defining of developmental stages of *Plasmodium* parasite and has been documented in open resource dataset [76]. *P. vivax* blood-stage-specific profile of expression using single-cell transcription allowed understanding of this neglected parasite in detail [77].

26.4 Proteomics in Malaria Biology

The central dogma of biology, i.e., DNA-RNA proteins, is the basis for understanding the in-depth functions of any organism. The field of omics dealing with the understanding of all proteins expressed in any organism at any given point of the life cycle has been termed as proteomics.

Major work in the field of *Plasmodium* proteomics came up in 2002 with two concurrent publications from Florens et al. and Lasonder et al. Both of these studies showed the presence of the different number of proteins expressed in *P. falciparum* parasite, i.e., nearly 2400 proteins across the proteome of different stages of the parasite have been identified by Florens et al. [78], whereas Lasonder et al. identified approximately 1289 proteins across the proteomes of gametocytes, gametes, and late asexual blood stages [79]. Out of the identified proteins across all stages and stage-specific proteins, only 30% highlighted the role of different biological functions in the parasite life cycle [78, 79]. The studies highlighted the abundance of proteins in the different stages of parasite according to the functions. As, invasion proteins were

found to be abundant in merozoites, whereas dynein proteins have been found to be high in sexual stages, it highlights the importance of proteomics studies which can reveal out important proteins required for the survival of a particular stage of the parasite and sex-specific proteins [55, 80].

Complete genome sequencing of the malaria parasite acted as a scaffold for the global analysis of the expression pattern of the parasite proteome. Out of 2400 proteins identified in the study by Florens et al., nearly half of the proteins were hypothetical, which are highly unique to the parasite as they lack any homology with any other protein in other species. Such a huge number of *Plasmodium*-specific proteins opened a new avenue for the identification of a safe and efficacious vaccine against malaria. A similar revelation was made by another concurrent study by Lasonder et al. who identified nearly 580 sex stage-specific proteins which can be very useful in finding out the important proteins required for the sexual stage development in the parasite [55].

Encouraged with the information provided by these two initial studies, stage-specific proteomics in *Plasmodium* has been carried out by a number of researchers in the last few years. As the researchers started looking into the details of the *Plasmodium* parasite proteome, some technical issues have been encountered. One of the major ones is the poor success rate of membrane protein identification and analyses. As these proteins have high hydrophobicity, they have been shown to be difficult to separate by two-dimensional gel electrophoresis and liquid chromatography. But as the most important proteins involved in the invasion process are either on the surface of merozoites or sporozoites, these proteins received tremendous attention. Thus, even after limitations in the current proteomics method, successful investigation of several membrane proteins of the parasite has been done successfully [81]. Detergent-resistant proteins identified using novel methods resulted in the identification of proteins associated with the organelles, membrane proteins, and protein part of exportome [81]. Small molecule-linked proteomics studies also helped in the identification of some of the important drug targets which have been involved directly in parasite invasion of erythrocytes, e.g., falcipain [82].

Several other important proteins including GPI-anchored proteins have been identified in the parasite using a combination of novel methods and proteomics [83]. Two of these proteins, merozoite surface protein 1 and merozoite surface protein 2, are known to make 2/3 of the total surface-exposed proteins on merozoites. The rest of the less abundant proteins have been identified using different strategies, one of those involved shaving of the exposed proteins on infected erythrocytes and merozoites using trypsinization [84].

The proteomics field has evolved in the last decade and has benefited malaria research in characterizing a subset of proteins that carry posttranslational modifications. Along with this, organelle-specific proteins, e.g., nuclear protein [85, 86], apicoplast proteome [87], ubiquitome [88], and selenocysteine-specific proteome, have been identified [89, 90]. Also, important surface-expressed proteins have been identified using advanced proteomics [91].

Proteomics also helped in identifying some of the important pathways employed by the parasite for its survival. As we know parasite stays within the erythrocyte, it

has developed a pathway to export protein to the surface of the infected RBC. Parasite has several strategies for hijacking host erythrocyte for several advantages to itself. One of such examples is the elaborate protein transport system that has been developed for carrying proteins to the surface of infected erythrocytes. This transport allow parasite to evade immune system by expressing its own proteins on the surface, as erythrocytes are the only non-antigen-presenting cells in blood. For a long time, it has not been known what is the mechanism of exporting these proteins to the surface of the infected erythrocytes. Using a combination of immune pulldowns and shotgun proteomics, a novel gene family has been identified which has been expressed along with rif gene product during trophozoite life stage [92].

Liver stage proteomics studies have also been carried out to understand this important stage of parasite life cycle. However due to the absence of in vitro models for this stage, not much has been done in this area [93]. However proteomics analysis of merosomes from *P. falciparum* has provided valuable information regarding link between blood stage and liver stage of the parasite [94]. Comparisons between schizont-infected cell antigens of *Plasmodium knowlesi* (the causative agent of monkey malaria) and *P. falciparum*'s PFEMP1 antigen extracellular domain have been carried out [95] and revealed that there is high homology between these important domains and the use of rhesus monkeys' model validates for the study of this family of surface proteins. Another study established the expression of var genes in the sporozoite stages [78]. Comparative proteomics and functional analysis helped in identifying role of "osmiophilic bodies" (electron-dense secretory organelles of the female gametocytes which discharge their contents during gamete formation) [96].

Proteomics helped in identifying the proteins which have been expressed at a particular stage of the parasite life cycle and also helped in identifying a vast number of genes exclusively expressed in the parasite. Recent studies have focused on the influence of drugs such as artemisinin on the expression of specific sets of proteins and pathways [94, 97]. Quantitative proteomics studies have revealed that there is important role of posttranslational modification of surface-exposed proteins during switching of invasion pathways utilized by the parasite [98]. Further these studies also explored various other posttranslational modifications in the parasite as arginine and methylation of various proteins [99].

Overall proteomics and its tools have widely been employed currently to decipher some of the long-awaited mysteries of parasite. These studies will not only allow us to understand pathogen well but also will be an asset for development of newer strategies for therapy and prevention of malaria (Fig. 26.1).

26.5 Metabolomics in Malaria Biology

Metabolomics has led to biological research into a new and unexplored territory. With the help of metabolomics, scientists and researchers have been able to bring together strong biochemical, mathematical, statistical, and bioinformatic procedures which further examine and analyze the ambiguities within disease area of wider



Fig. 26.1 Time line of the major *Plasmodium* spp. sequencing projects, genome, transcriptome, and proteome sequencing projects

biology. The usage of metabolomics to learn about malaria parasites characterizes a significant advancement in gaining a better understanding on parasite biology and disease etiology [100].

Furthermore, metabolomics is an important technology that has enabled the understanding of metabolism and its components in detail. Metabolism is key to the phenotype of an organism as an important step of functional genomics to understand changes in the metabolites during normal and perturbed conditions. As parasite undergoes various changes during its life cycle and exposure to drugs, metabolomics studies become of keen interest to understand important steps that need to be targeted for better therapeutics.

As per definition metabolomics is an approach of quantitative and qualitative analysis of a set of small molecules present in any biological sample under a given set of physiological conditions [101]. The approach to understanding metabolome is very challenging, and the complexity of the system needs constant modifications in the existing analytical platforms, including mass spectrometry NMR and chromatography techniques. Looking at the *Plasmodium* complexity, metabolomics studies proved to be very elucidating in parasite biology. There are very few in-depth metabolomics studies that have been performed in malaria biology; however, various researchers have speculated on the utility of system biology approaches in understanding the metabolic analysis of the parasite [102–104].

There is an urgent need for in-depth metabolomics analyses of various parasite-related topics. To start with, drug and vaccine target identification is a major area of research, in which the metabolomics approach can be helpful. The complex life cycle of the parasite has posed various challenges which are tried to be countered using genomics and proteomics approaches. These approaches helped in investigating expression profiles and to function prediction for the number of genes [3, 105]. But all these studies are conclusive in the identification of gene function. Similarly, proteomics studies have identified a large number of proteins which have not been assigned with functions. In such a scenario, the acquisition of metabolomics data on parasite biochemical process can be of great help in supporting existing datasets and provides us with a clear picture.

One of the fields which have been explored well with the metabolomics approach in the parasite is lipid metabolism. Lipid metabolism is a suitable pathway to be targeted for drug target identification as it lacks many homologies with the host proteins [106–109]. Major changes in the lipid profile of the host has been reported in various protozoan infection [110] including malaria, where lipid metabolism is altered during the severe malaria infections [111].

Metabolite profiling is an important aspect of metabolomics as it provides biochemical pieces of evidence for the functional role of putative enzymes and also facilitates annotation of new metabolic pathways in the parasite [112, 113]. Drug screens based on metabolomics have helped in the identification of targets for potent antimalarials from Malaria Box [114, 115] and highlighted some of the essential unknown pathways of the parasite.

In Parasite, the level of intracellular metabolite pools from different parasite stages and variation in the extracellular levels of metabolites is been utilized to

infer the operation of cultured parasite stages and specific pathways [116]. Kinetic flux profiling has helped in understanding the role of pyruvate dehydrogenase complex and showed that it plays no major role in acetyl CoA synthesis, tricarboxylic acid metabolism, or fatty acid synthesis in blood-stage parasites [116].

From a metabolomics perspective, it is also important to understand that five main species of *Plasmodium* cause malaria in the human body. Moreover, it is the responses of human immune system and the interactions between human and parasites that occur during the infection which lead to complications of malaria (*P. falciparum*, *P. malariae*, *P. ovale*, and *P. vivax*). Genome studies have advanced to the extent that they now help understand the genetic bases of host susceptibility and pathogen life cycle. Metabolomics is unique because it reveals thousands of metabolites which show the function of parasites, the host species, and the exposure to the environment. Metabolomics is key because when coupled with other systems of biology it can help aid with easier and better diagnosis, treatment, and ultimately the eradication of malaria [117].

26.6 Glycomics in *Plasmodium* Biology

Glycomics is a recent development in the fields of omics which deal with identifying and characterizing polysaccharides or sugar structures, of which “glycome” is made up of [118]. Like other recent advances in the field, glycomics also faces some challenges, e.g., nonlinearity and posttranslational modifications [118, 119].

In *Plasmodium*, structures related to glycans are glycosylphosphatidylinositol (GPI) anchors [120, 121]. Recently few N-glycans made up of one or two GlcNac residues [122, 123] have also been considered to be part of this list. The structures of *Plasmodium* GPI are well defined; however, the presence of glycosylation and the level of N-glycosylation is still controversial [124, 125].

Glycans play an important role in malaria biology as they are involved in glycan-based interaction during host-pathogen interactions. *Plasmodium* relies heavily on glycan-based interaction as it is true with other pathogens [126, 127]. It is not only the invasion process of *Plasmodium* which highly relies on glycan interactions, but also the severity of the infection has been governed by these glycans. Infected erythrocytes and glycosaminoglycan interactions on the endothelial lining of vessels reduce splenic clearance and lead to severe cases, coma, and organ failure [128]. Another important feature of *Plasmodium* infection, i.e., “rosetting,” has also been influenced by glycan interactions with surface-exposed PfEMP1 [129–131]. Rosetting further leads to blockages of the blood vessels and impedes circulation leading to severe form of malaria [132, 133]. Glycomics can be very helpful in understanding these glycans important for interaction among these important proteins and carbohydrates.

O-GlcNacylated protein is one of the examples of PTMs that have been observed in another organism including *Plasmodium*. Several proteins have been identified in the parasite having these modifications [134] which have an important role to play in the biology of the parasite.

Improving tools of glycomics and increasing importance of glycans in host-pathogen interactions have opened a new avenue for field of glycomics in malaria biology, which will help us in understanding the pathogen in a better way.

26.7 Kinomics in Malaria Biology

Phosphorylation of proteins is known to play an essential and important role in signaling pathways, which ultimately lead to the phenotype of the organism. Kinases and dephosphorylates are the enzymes carrying out this function widely in any given organism. Reverse genetic approaches have deciphered a number of these enzymes, i.e., kinases which are essential for asexual stages of the parasite. However, phosphoproteomics has been key for identifying substrates and the pathways that they regulate [135]. Further studies focused on stage-specific expressions of phosphorylated proteins to understand the complexity of the life cycle [136–138]. Quantitative phosphoproteomics studies highlighted the important role of host kinase modification during different important steps of *Plasmodium* life cycle, i.e., egress [139, 140].

Major events of parasite have been governed by phosphorylation status of various proteins in the parasite, and hence the tools of kinomics will be widely employed in the coming future to find out key target proteins responsible for governing these events in the parasite.

26.8 Omics and the Way Forward

26.8.1 Drug Target Identification

In light of the drug resistance and the vast spread of malaria infection, the development of new drug targets and effective vaccine candidates is of urgent requirement. With advances in omics technology, tremendous information about *Plasmodium* genome proteome and metabolome has been deciphered. Genome sequencing of *P. falciparum* in 2002 [3] and *P. vivax* in 2008 [141] leads to the fasten discovery of drug targets. New databases devoted to information regarding *Plasmodium* have been developed which helped the researchers to find new targets. An example of such databases is web-based PlasmoDB (<http://www.plasmodb.org>) [142]. This database has been operated under the umbrella of Eukaryotic Pathogen Genomics Database (<http://EuPathDB.org>) on contract from the US National Institute of Allergy and Infectious Diseases (NIAID).

Whole-genome sequencing opened new avenues for the important metabolic pathways and associated essential parasite-specific important genes [25, 143–147]. One such example comes from the identification of small molecule inhibitor LY411,575 which targets signal peptide peptidase (putative protease component) of ER-associated degradation pathway (ERAD) [148]. In this study, a bioinformatic-based orthologue detection approach identified a highly simplified ERAD pathway

in *P. falciparum* as compared to mammalian cells, which are evident in the effectiveness of small molecule inhibitors of ERAD system. Spiroindolone class (spiroetrahydro-b-carbolines) inhibitor of protein synthesis is also a good example of such small molecules' identification with the same approach [146]. Similarly, inhibitor against histone methyltransferase compounds BIX-01294 and TM2-115 have been identified [149].

Understanding changes in the gene expression profile due to chemical changes or drugs can shed light on the mechanism of drug action and can be very helpful in rational drug development. In *Plasmodium*, widely used drugs such as artemisinin, chloroquine, quinine, antifolate, and doxycycline have been used to understand transcriptome [64, 150, 151] and proteome [152, 153] of the parasite under the effect of these drugs. Data from these studies suggested alterations in a lot of functionally important groups of genes of low amplitude, which could not have been detected using a general approach. For example, Proteomics studies with doxycycline established Mitochondria and Apicoplast as site of important pathways which can be targeted with small molecule inhibitors [152].

26.8.2 Vaccine Targets and OMICS

The vaccine for malaria is the most challenging task for researchers. Complex multistage life cycle, differential expressions of different proteins, the role of proteins in different regulations antigenic variation, and extensive genetic diversity of critical target epitopes are the factors which make the task more challenging. Currently, knowledge about the genome and proteome of the parasite has aided in the development of new vaccine strategies. Due to advance genome sequencing platforms, scans of polymorphisms throughout the genome enable identification of genes encoding important targets of immunity [143, 154–156].

Transcriptomic data can also be used to reveal unknown genes which can be good drug targets for generating protective immunity [51, 52, 56, 105, 157–159]. The transcriptomic analysis helped in the identification of some known genes which have been missed using previous technologies. In one of such study, 262 orfs, not identified earlier, have been identified using transcriptomic expression profiles. Another important information came from a study exploring *P. falciparum* transcriptome exploring genes for early gametocytogenesis, an important step in parasite transmission. Data from the study pointed out toward continuous production of gametocytes during asexual growth, suggesting continued transmission in the field during human infections. This leads to the importance of methods for malaria control directed toward this important stage of the parasite [160].

Advances in high-throughput sequencing allowing quantitative assessment of RNA transcripts expressed (RNA Seq) in a cell allowed a new information about *P. falciparum* transcriptome, leading to identification of novel transcripts, low abundance transcript, and splicing variants [54, 161–163].

26.9 Omics to Biology of the Parasite

Omics is a combination of various “omes” which should be understood and analyzed in conjunction to make a correct understanding of parasite biology. *P. falciparum* biology in light of functional omics has enlightened researchers to great extent about genes specifically associated with parasite under specific conditions. Transcriptome analysis of parasite under high density revealed genes associated with stress-related and cell death [164]. Combination of transcriptome and proteome highlighted large-scale translational repression as a hallmark of female gametocyte biology [55]. Plant-like α -linolenic acid pathway (ALA) has been identified in the parasite using metabolite profiling [112]. Similarly, the multi-omics approach has helped in understanding the role of posttranslational modification of proteins in the invasion process [165].

One of the major success stories revealed due to the multi-omics approach is knowledge about gametocytogenesis, the important process of conversion of the asexual parasite to the sexual stage. A long search for the regulators responsible for this conversion was carried out, which finally started with homology-based identification of transcription factors. AP2 family of transcription regulators from plant and algae was the first reported regulator of the gametocytogenesis process [166]. The discovery has been corroborated with transcriptional studies which showed expression levels of these regulators in all the stages [166]. Further transcriptomic studies identified AP2-G as the key regulator along with other markers for gametocytogenesis [63, 65]. Single-cell RNA analysis further enhanced our knowledge in the area by identifying AP2-G+, and analysis of sexually committed schizonts (AP2-G+) has shown promising results in characterizing the transcriptional program of gametocytogenesis [74]. Sexual conversion of the parasite into gametocyte can follow two alternative routes, i.e., either in the same cycle through AP2G+ or in the subsequent cycle, with specific transcription signatures [167].

26.10 Summary

Malaria, being one of the most common infectious diseases in today's era, needs special medication and drug target due to very few numbers of drugs and development of resistance in parasite. In the modern world, research is more concentrated around “omics” biology. The “omics” era includes genomics, transcriptomics, proteomics, metabolomics, glycomics, and kinomics. These studies have aided in a better understanding of malarial biology and helped fasten the finding of drugs and vaccine targets. Genetic diversity has been explored with the help of various developments in the genomics approach. These findings will enable the development of various therapies and drug resistance monitoring. Transcriptomics and proteomics together can help in the identification of expression patterns at different parasite stages. Such information can help in identifying better drug targets. Moreover, metabolomics studies are of great interest as they are essential to understand multiple phenotypic changes during the parasite life cycle and drug exposure. Another major

aspect of the understanding host-pathogen interaction is glycans, which has brought new opportunities in the field of glycomics for a better understanding of pathogen. In the near future, kinomics will be extensively used to recognize the main target proteins involved in controlling the phosphorylation of proteins in the parasite.

The present goal is to refine the available data to obtain useful data, like the discovery of novel drug targets and candidates for vaccines. For the high-throughput detection of potential candidates, many bioinformatics methods and emerging technologies have been utilized. The omics techniques, combined with other systems of biology, can help with faster and improved diagnosis, treatment, and potentially the eradication of malaria.

References

1. Hardison RC (2003) Comparative genomics. *PLoS Biol* 1(2):E58. <https://doi.org/10.1371/journal.pbio.0000058>
2. Butler D (1997) Funding assured for international malaria sequencing project. *Nature* 388(6644):701. <https://doi.org/10.1038/41826>
3. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW et al (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419(6906):498–511. <https://doi.org/10.1038/nature01097>
4. Bright AT, Tewhey R, Abeles S, Chuquiyauri R, Llanos-Cuentas A, Ferreira MU et al (2012) Whole genome sequencing analysis of *Plasmodium vivax* using whole genome capture. *BMC Genomics* 13:262. <https://doi.org/10.1186/1471-2164-13-262>
5. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Perlea M, Silva JC et al (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* 419(6906):512–519. <https://doi.org/10.1038/nature01099>
6. Pain A, Bohme U, Berry AE, Mungall K, Finn RD, Jackson AP et al (2008) The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* 455(7214):799–803. <https://doi.org/10.1038/nature07306>
7. Tachibana S, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N et al (2012) *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. *Nat Genet* 44(9):1051–1055. <https://doi.org/10.1038/ng.2375>
8. Rutledge GG, Marr I, Huang GKL, Auburn S, Marfurt J, Sanders M et al (2017) Genomic characterization of recrudescence *Plasmodium malariae* after treatment with Artemether/Lumefantrine. *Emerg Infect Dis* 23(8):1300–1307. <https://doi.org/10.3201/eid2308.161582>
9. Otto TD, Rayner JC, Bohme U, Pain A, Spottiswoode N, Sanders M et al (2014) Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat Commun* 5:4754. <https://doi.org/10.1038/ncomms5754>
10. Otto TD, Gilbert A, Crellen T, Bohme U, Arnathau C, Sanders M et al (2018) Genomes of all known members of a *Plasmodium* subgenus reveal paths to virulent human malaria. *Nat Microbiol* 3(6):687–697. <https://doi.org/10.1038/s41564-018-0162-2>
11. Escalante AA, Ayala FJ (1994) Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences. *Proc Natl Acad Sci U S A* 91(24):11373–11377. <https://doi.org/10.1073/pnas.91.24.11373>
12. Perkins SL, Sarkar IN, Carter R (2007) The phylogeny of rodent malaria parasites: simultaneous analysis across three genomes. *Infect Genet Evol* 7(1):74–83. <https://doi.org/10.1016/j.meegid.2006.04.005>
13. Loy DE, Liu W, Li Y, Learn GH, Plenderleith LJ, Sundararaman SA et al (2017) Out of Africa: origins and evolution of the human malaria parasites *Plasmodium falciparum* and *Plasmodium vivax*. *Int J Parasitol* 47(2–3):87–97. <https://doi.org/10.1016/j.ijpara.2016.05.008>

14. Talundzic E, Ravishankar S, Kelley J, Patel D, Plucinski M, Schmedes S et al (2018) Next-generation sequencing and bioinformatics protocol for malaria drug resistance marker surveillance. *Antimicrob Agents Chemother* 62(4). <https://doi.org/10.1128/AAC.02474-17>
15. de Oliveira TC, Corder RM, Early A, Rodrigues PT, Ladeia-Andrade S, Alves JMP et al (2020) Population genomics reveals the expansion of highly inbred *Plasmodium vivax* lineages in the main malaria hotspot of Brazil. *PLoS Negl Trop Dis* 14(10):e0008808. <https://doi.org/10.1371/journal.pntd.0008808>
16. Hupalo DN, Luo Z, Melnikov A, Sutton PL, Rogov P, Escalante A et al (2016) Population genomics identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. *Nat Genet* 48(8):953–958. <https://doi.org/10.1038/ng.3588>
17. Ocholla H, Preston MD, Mipando M, Jensen AT, Campino S, MacInnis B et al (2014) Whole-genome scans provide evidence of adaptive evolution in Malawian *Plasmodium falciparum* isolates. *J Infect Dis* 210(12):1991–2000. <https://doi.org/10.1093/infdis/jiu349>
18. Miotto O, Almagro-Garcia J, Manske M, Macinnis B, Campino S, Rockett KA et al (2013) Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia. *Nat Genet* 45(6):648–655. <https://doi.org/10.1038/ng.2624>
19. Su X, Ferdig MT, Huang Y, Huynh CQ, Liu A, You J et al (1999) A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* 286(5443):1351–1353. <https://doi.org/10.1126/science.286.5443.1351>
20. Aravind L, Iyer LM, Welles TE, Miller LH (2003) *Plasmodium* biology: genomic gleanings. *Cell* 115(7):771–785. [https://doi.org/10.1016/s0092-8674\(03\)01023-7](https://doi.org/10.1016/s0092-8674(03)01023-7)
21. Su X, Hayton K, Welles TE (2007) Genetic linkage and association analyses for trait mapping in *Plasmodium falciparum*. *Nat Rev Genet* 8(7):497–506. <https://doi.org/10.1038/nrg2126>
22. Kooij TW, Carlton JM, Bidwell SL, Hall N, Ramesar J, Janse CJ et al (2005) A *Plasmodium* whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes. *PLoS Pathog* 1(4):e44. <https://doi.org/10.1371/journal.ppat.0010044>
23. Carlton J, Silva J, Hall N (2005) The genome of model malaria parasites, and comparative genomics. *Curr Issues Mol Biol* 7(1):23–37
24. Baum J, Thomas AW, Conway DJ (2003) Evidence for diversifying selection on erythrocyte-binding antigens of *Plasmodium falciparum* and *P. vivax*. *Genetics* 163(4):1327–1336
25. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M et al (2005) A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science* 307(5706):82–86. <https://doi.org/10.1126/science.1103717>
26. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, Feldman M et al (1995) Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* 82(1):77–87. [https://doi.org/10.1016/0092-8674\(95\)90054-3](https://doi.org/10.1016/0092-8674(95)90054-3)
27. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS et al (1995) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* 82(1):89–100. [https://doi.org/10.1016/0092-8674\(95\)90055-1](https://doi.org/10.1016/0092-8674(95)90055-1)
28. Smith T, Charlwood JD, Takken W, Tanner M, Spiegelhalter DJ (1995) Mapping the densities of malaria vectors within a single village. *Acta Trop* 59(1):1–18. [https://doi.org/10.1016/0001-706x\(94\)00082-c](https://doi.org/10.1016/0001-706x(94)00082-c)
29. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F et al (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* 407(6807):1018–1022. <https://doi.org/10.1038/35039531>
30. Taylor HM, Kyes SA, Newbold CI (2000) Var gene diversity in *Plasmodium falciparum* is generated by frequent recombination events. *Mol Biochem Parasitol* 110(2):391–397. [https://doi.org/10.1016/s0166-6851\(00\)00286-3](https://doi.org/10.1016/s0166-6851(00)00286-3)

31. Voss TS, Healer J, Marty AJ, Duffy MF, Thompson JK, Beeson JG et al (2006) A var gene promoter controls allelic exclusion of virulence genes in *Plasmodium falciparum* malaria. *Nature* 439(7079):1004–1008. <https://doi.org/10.1038/nature04407>
32. Scherff A, Khanavkar B, Ostendorf U, Phillipou S, Ewig S (2006) Central airway obstruction in ulcerative colitis—a rare extraintestinal manifestation. *Pneumologie* 60(10):607–610. <https://doi.org/10.1055/s-2006-944249>
33. Robinson BA, Welch TL, Smith JD (2003) Widespread functional specialization of *Plasmodium falciparum* erythrocyte membrane protein 1 family members to bind CD36 analysed across a parasite genome. *Mol Microbiol* 47(5):1265–1278. <https://doi.org/10.1046/j.1365-2958.2003.03378.x>
34. Trimnell AR, Kraemer SM, Mukherjee S, Phippard DJ, Janes JH, Flamoe E et al (2006) Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria. *Mol Biochem Parasitol* 148(2):169–180. <https://doi.org/10.1016/j.molbiopara.2006.03.012>
35. Kirchgatter K, Del Portillo HA (2005) Clinical and molecular aspects of severe malaria. *An Acad Bras Cienc* 77(3):455–475. <https://doi.org/10.1590/s0001-37652005000300008>
36. Kaestli M, Cockburn IA, Cortes A, Baea K, Rowe JA, Beck HP (2006) Virulence of malaria is associated with differential expression of *Plasmodium falciparum* var gene subgroups in a case-control study. *J Infect Dis* 193(11):1567–1574. <https://doi.org/10.1086/503776>
37. Bull PC, Berriman M, Kyes S, Quail MA, Hall N, Kortok MM et al (2005) *Plasmodium falciparum* variant surface antigen expression patterns during malaria. *PLoS Pathog* 1(3):e26. <https://doi.org/10.1371/journal.ppat.0010026>
38. Duffy MF, Caragounis A, Noviyanti R, Kyriacou HM, Choong EK, Boysen K et al (2006) Transcribed var genes associated with placental malaria in Malawian women. *Infect Immun* 74(8):4875–4883. <https://doi.org/10.1128/IAI.01978-05>
39. Semblat JP, Raza A, Kyes SA, Rowe JA (2006) Identification of *Plasmodium falciparum* var1CSA and var2CSA domains that bind IgM natural antibodies. *Mol Biochem Parasitol* 146(2):192–197. <https://doi.org/10.1016/j.molbiopara.2005.12.007>
40. Amambua-Ngwa A, Jeffries D, Amato R, Worwui A, Karim M, Ceesay S et al (2018) Consistent signatures of selection from genomic analysis of pairs of temporal and spatial *Plasmodium falciparum* populations from the Gambia. *Sci Rep* 8(1):9687. <https://doi.org/10.1038/s41598-018-28017-5>
41. Duffy CW, Amambua-Ngwa A, Ahouidi AD, Diakite M, Awandare GA, Ba H et al (2018) Multi-population genomic analysis of malaria parasites indicates local selection and differentiation at the *gdv1* locus regulating sexual development. *Sci Rep* 8(1):15763. <https://doi.org/10.1038/s41598-018-34078-3>
42. Demas AR, Sharma AI, Wong W, Early AM, Redmond S, Bopp S et al (2018) Mutations in *Plasmodium falciparum* actin-binding protein coronin confer reduced artemisinin susceptibility. *Proc Natl Acad Sci U S A* 115(50):12799–12804. <https://doi.org/10.1073/pnas.1812317115>
43. Bankole BE, Kayode AT, Nosamiefan IO, Eromon P, Baniecki ML, Daniels RF et al (2018) Characterization of *Plasmodium falciparum* structure in Nigeria with malaria SNPs barcode. *Malar J* 17(1):472. <https://doi.org/10.1186/s12936-018-2623-8>
44. Bei AK, Niang M, Deme AB, Daniels RF, Sarr FD, Sokhna C et al (2018) Dramatic changes in malaria population genetic complexity in Dielmo and Ndiop, Senegal, revealed using genomic surveillance. *J Infect Dis* 217(4):622–627. <https://doi.org/10.1093/infdis/jix580>
45. Chang HH, Worby CJ, Yeka A, Nankabirwa J, Kanya MR, Staedke SG et al (2017) THE REAL McCOIL: a method for THE concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. *PLoS Comput Biol* 13(1):e1005348. <https://doi.org/10.1371/journal.pcbi.1005348>
46. Trevino SG, Nkhoma SC, Nair S, Daniel BJ, Moncada K, Khoswe S et al (2017) High-resolution single-cell sequencing of malaria parasites. *Genome Biol Evol* 9(12):3373–3383. <https://doi.org/10.1093/gbe/evx256>

47. Nair S, Nkhoma SC, Serre D, Zimmerman PA, Gorena K, Daniel BJ et al (2014) Single-cell genomics for dissection of complex malaria infections. *Genome Res* 24(6):1028–1038. <https://doi.org/10.1101/gr.168286.113>
48. Nkhoma SC, Banda RL, Khoswe S, Dzoole-Mwale TJ, Ward SA (2018) Intra-host dynamics of co-infecting parasite genotypes in asymptomatic malaria patients. *Infect Genet Evol* 65:414–424. <https://doi.org/10.1016/j.meegid.2018.08.018>
49. Evans GA (2000) Designer science and the “omic” revolution. *Nat Biotechnol* 18(2):127. <https://doi.org/10.1038/72480>
50. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL (2003) Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biol* 4(2):R9. <https://doi.org/10.1186/gb-2003-4-2-r9>
51. Young JA, Fivelman QL, Blair PL, de la Vega P, Le Roch KG, Zhou Y et al (2005) The *Plasmodium falciparum* sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol* 143(1):67–79. <https://doi.org/10.1016/j.molbiopara.2005.05.007>
52. Siau A, Silvie O, Franetich JF, Yalaoui S, Marinach C, Hannoun L et al (2008) Temperature shift and host cell contact up-regulate sporozoite expression of *Plasmodium falciparum* genes involved in hepatocyte infection. *PLoS Pathog* 4(8):e1000121. <https://doi.org/10.1371/journal.ppat.1000121>
53. Lopez-Barragan MJ, Lemieux J, Quinones M, Williamson KC, Molina-Cruz A, Cui K et al (2011) Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. *BMC Genomics* 12:587. <https://doi.org/10.1186/1471-2164-12-587>
54. Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, Bohme U et al (2010) New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Mol Microbiol* 76(1):12–24. <https://doi.org/10.1111/j.1365-2958.2009.07026.x>
55. Lasonder E, Rijpma SR, van Schaijk BC, Hoeijmakers WA, Kensch PR, Gresnigt MS et al (2016) Integrated transcriptomic and proteomic analyses of *P. falciparum* gametocytes: molecular insight into sex-specific processes and translational repression. *Nucleic Acids Res* 44(13):6087–6101. <https://doi.org/10.1093/nar/gkw536>
56. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD et al (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 301(5639):1503–1508. <https://doi.org/10.1126/science.1087025>
57. Horrocks P, Decherig K, Lanzer M (1998) Control of gene expression in *Plasmodium falciparum*. *Mol Biochem Parasitol* 95(2):171–181. [https://doi.org/10.1016/s0166-6851\(98\)00110-8](https://doi.org/10.1016/s0166-6851(98)00110-8)
58. Daily JP, Le Roch KG, Sarr O, Fang X, Zhou Y, Ndir O et al (2004) In vivo transcriptional profiling of *Plasmodium falciparum*. *Malar J* 3:30. <https://doi.org/10.1186/1475-2875-3-30>
59. Daily JP, Le Roch KG, Sarr O, Ndiaye D, Lukens A, Zhou Y et al (2005) In vivo transcriptome of *Plasmodium falciparum* reveals overexpression of transcripts that encode surface proteins. *J Infect Dis* 191(7):1196–1203. <https://doi.org/10.1086/428289>
60. Barbour AG, Dai Q, Restrepo BI, Stoenner HG, Frank SA (2006) Pathogen escape from host immunity by a genome program for antigenic variation. *Proc Natl Acad Sci U S A* 103(48):18290–18295. <https://doi.org/10.1073/pnas.0605302103>
61. Frank SA, Barbour AG (2006) Within-host dynamics of antigenic variation. *Infect Genet Evol* 6(2):141–146. <https://doi.org/10.1016/j.meegid.2004.10.005>
62. Dzikowski R, Templeton TJ, Deitsch K (2006) Variant antigen gene expression in malaria. *Cell Microbiol* 8(9):1371–1381. <https://doi.org/10.1111/j.1462-5822.2006.00760.x>
63. Reid AJ, Talman AM, Bennett HM, Gomes AR, Sanders MJ, Illingworth CJR et al (2018) Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *elife* 7. <https://doi.org/10.7554/eLife.33105>
64. Hu G, Cabrera A, Kono M, Mok S, Chaa BK, Haase S et al (2010) Transcriptional profiling of growth perturbations of the human malaria parasite *Plasmodium falciparum*. *Nat Biotechnol* 28(1):91–98. <https://doi.org/10.1038/nbt.1597>

65. Rovira-Graells N, Gupta AP, Planet E, Crowley VM, Mok S, Ribas de Pouplana L et al (2012) Transcriptional variation in the malaria parasite *Plasmodium falciparum*. *Genome Res* 22 (5):925–938. <https://doi.org/10.1101/gr.129692.111>
66. Chakrabarti K, Pearson M, Grate L, Sterne-Weiler T, Deans J, Donohue JP et al (2007) Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *RNA* 13(11):1923–1939. <https://doi.org/10.1261/rna.751807>
67. Broadbent KM, Broadbent JC, Ribacke U, Wirth D, Rinn JL, Sabeti PC (2015) Strand-specific RNA sequencing in *Plasmodium falciparum* malaria identifies developmentally regulated long non-coding RNA and circular RNA. *BMC Genomics* 16:454. <https://doi.org/10.1186/s12864-015-1603-4>
68. Filarsky M, Frasncka SA, Niederwieser I, Brancucci NMB, Carrington E, Carrio E et al (2018) GDVI induces sexual commitment of malaria parasites by antagonizing HP1-dependent gene silencing. *Science* 359(6381):1259–1263. <https://doi.org/10.1126/science.aan6042>
69. Jing Q, Cao L, Zhang L, Cheng X, Gilbert N, Dai X et al (2018) *Plasmodium falciparum* var gene is activated by its antisense long noncoding RNA. *Front Microbiol* 9:3117. <https://doi.org/10.3389/fmicb.2018.03117>
70. Xue X, Zhang Q, Huang Y, Feng L, Pan W (2008) No miRNA were found in *Plasmodium* and the ones identified in erythrocytes could not be correlated with infection. *Malar J* 7:47. <https://doi.org/10.1186/1475-2875-7-47>
71. Rathjen T, Nicol C, McConkey G, Dalmay T (2006) Analysis of short RNAs in the malaria parasite and its red blood cell host. *FEBS Lett* 580(22):5185–5188. <https://doi.org/10.1016/j.febslet.2006.08.063>
72. Baum J, Papenfuss AT, Mair GR, Janse CJ, Vlachou D, Waters AP et al (2009) Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites. *Nucleic Acids Res* 37(11):3788–3798. <https://doi.org/10.1093/nar/gkp239>
73. Achieng AO, Hengartner NW, Raballah E, Cheng Q, Anyona SB, Lauve N et al (2019) Integrated OMICS platforms identify LAIR1 genetic variants as novel predictors of cross-sectional and longitudinal susceptibility to severe malaria and all-cause mortality in Kenyan children. *EBioMedicine* 45:290–302. <https://doi.org/10.1016/j.ebiom.2019.06.043>
74. Poran A, Notzel C, Aly O, Mencia-Trinchant N, Harris CT, Guzman ML et al (2017) Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. *Nature* 551(7678):95–99. <https://doi.org/10.1038/nature24280>
75. Brancucci NMB, De Niz M, Straub TJ, Ravel D, Sollelis L, Birren BW et al (2018) Probing *Plasmodium falciparum* sexual commitment at the single-cell level. *Wellcome Open Res* 3:70. <https://doi.org/10.12688/wellcomeopenres.14645.4>
76. Howick VM, Russell AJC, Andrews T, Heaton H, Reid AJ, Natarajan K et al The malaria cell atlas: single parasite transcriptomes across the complete *Plasmodium* life cycle. *Science* 2019 (6455):365. <https://doi.org/10.1126/science.aaw2619>
77. Sa JM, Cannon MV, Caleon RL, Wellems TE, Serre D (2020) Single-cell transcription analysis of *Plasmodium vivax* blood-stage parasites identifies stage- and species-specific profiles of expression. *PLoS Biol* 18(5):e3000711. <https://doi.org/10.1371/journal.pbio.3000711>
78. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD et al (2002) A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* 419(6906):520–526. <https://doi.org/10.1038/nature01107>
79. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW et al (2002) Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* 419(6906):537–542. <https://doi.org/10.1038/nature01111>
80. Meerstein-Kessel L, Andolina C, Carrio E, Mahamar A, Sawa P, Diawara H et al (2018) A multiplex assay for the sensitive detection and quantification of male and female *Plasmodium falciparum* gametocytes. *Malar J* 17(1):441. <https://doi.org/10.1186/s12936-018-2584-y>

81. Sanders PR, Gilson PR, Cantin GT, Greenbaum DC, Nebl T, Carucci DJ et al (2005) Distinct protein classes including novel merozoite surface antigens in raft-like membranes of *Plasmodium falciparum*. *J Biol Chem* 280(48):40169–40176. <https://doi.org/10.1074/jbc.M509631200>
82. Greenbaum DC, Baruch A, Grainger M, Bozdech Z, Medzihradsky KF, Engel J et al (2002) A role for the protease falcipain 1 in host cell invasion by the human malaria parasite. *Science* 298(5600):2002–2006. <https://doi.org/10.1126/science.1077426>
83. Gilson PR, Nebl T, Vukcevic D, Moritz RL, Sargeant T, Speed TP et al (2006) Identification and stoichiometry of glycosylphosphatidylinositol-anchored membrane proteins of the human malaria parasite *Plasmodium falciparum*. *Mol Cell Proteomics* 5(7):1286–1299. <https://doi.org/10.1074/mcp.M600035-MCP200>
84. Winter G, Kawai S, Haeggstrom M, Kaneko O, von Euler A, Kawazu S et al (2005) SURFIN is a polymorphic antigen expressed on *Plasmodium falciparum* merozoites and infected erythrocytes. *J Exp Med* 201(11):1853–1863. <https://doi.org/10.1084/jem.20041392>
85. Briquet S, Ourimi A, Pionneau C, Bernardes J, Carbone A, Chardonnet S et al (2018) Identification of *Plasmodium falciparum* nuclear proteins by mass spectrometry and proposed protein annotation. *PLoS One* 13(10):e0205596. <https://doi.org/10.1371/journal.pone.0205596>
86. Oehring SC, Woodcroft BJ, Moes S, Wetzel J, Dietz O, Pulfer A et al (2012) Organellar proteomics reveals hundreds of novel nuclear proteins in the malaria parasite *Plasmodium falciparum*. *Genome Biol* 13(11):R108. <https://doi.org/10.1186/gb-2012-13-11-r108>
87. Boucher MJ, Ghosh S, Zhang L, Lal A, Jang SW, Ju A et al (2018) Integrative proteomics and bioinformatic prediction enable a high-confidence apicoplast proteome in malaria parasites. *PLoS Biol* 16(9):e2005895. <https://doi.org/10.1371/journal.pbio.2005895>
88. Ponts N, Saraf A, Chung DW, Harris A, Prudhomme J, Washburn MP et al (2011) Unraveling the ubiquitome of the human malaria parasite. *J Biol Chem* 286(46):40320–40330. <https://doi.org/10.1074/jbc.M111.238790>
89. Lobanov AV, Delgado C, Rahlfs S, Novoselov SV, Kryukov GV, Gromer S et al (2006) The *Plasmodium* selenoproteome. *Nucleic Acids Res* 34(2):496–505. <https://doi.org/10.1093/nar/gkj450>
90. Roseler A, Prieto JH, Iozef R, Hecker B, Schirmer RH, Kulzer S et al (2012) Insight into the selenoproteome of the malaria parasite *Plasmodium falciparum*. *Antioxid Redox Signal* 17(4):534–543. <https://doi.org/10.1089/ars.2011.4276>
91. Nilsson Bark SK, Ahmad R, Dantzer K, Lukens AK, De Niz M, Szucs MJ et al (2018) Quantitative proteomic profiling reveals novel *Plasmodium falciparum* surface antigens and possible vaccine candidates. *Mol Cell Proteomics* 17(1):43–60. <https://doi.org/10.1074/mcp.RA117.000076>
92. Sam-Yellowe TY, Florens L, Johnson JR, Wang T, Drazba JA, Le Roch KG et al (2004) A *Plasmodium* gene family encoding Maurer's cleft membrane proteins: structural properties and expression profiling. *Genome Res* 14(6):1052–1059. <https://doi.org/10.1101/gr.2126104>
93. Sims PF, Hyde JE (2006) Proteomics of the human malaria parasite *Plasmodium falciparum*. *Expert Rev Proteomics* 3(1):87–95. <https://doi.org/10.1586/14789450.3.1.87>
94. Shears MJ, Sekhar Nirujogi R, Swearingen KE, Renuse S, Mishra S, Jaipal Reddy P et al (2019) Proteomic analysis of *Plasmodium* Merosomes: the link between liver and blood stages in malaria. *J Proteome Res* 18(9):3404–3418. <https://doi.org/10.1021/acs.jproteome.9b00324>
95. Korir CC, Galinski MR (2006) Proteomic studies of *Plasmodium knowlesi* SICA variant antigens demonstrate their relationship with *P. falciparum* EMP1. *Infect Genet Evol* 6(1):75–79. <https://doi.org/10.1016/j.meegid.2005.01.003>
96. Suarez-Cortes P, Sharma V, Bertuccini L, Costa G, Bannerman NL, Sannella AR et al (2016) Comparative proteomics and functional analysis reveal a role of *Plasmodium falciparum* Osmiophilic bodies in malaria parasite transmission. *Mol Cell Proteomics* 15(10):3243–3255. <https://doi.org/10.1074/mcp.M116.060681>

97. Graves PR, Kwiek JJ, Fadden P, Ray R, Hardeman K, Coley AM et al (2002) Discovery of novel targets of quinoline drugs in the human purine binding proteome. *Mol Pharmacol* 62 (6):1364–1372. <https://doi.org/10.1124/mol.62.6.1364>
98. Kuss C, Gan CS, Gunalan K, Bozdech Z, Sze SK, Preiser PR (2012) Quantitative proteomics reveals new insights into erythrocyte invasion by *Plasmodium falciparum*. *Mol Cell Proteomics* 11(2):M111 010645. <https://doi.org/10.1074/mcp.M111.010645>
99. Zeeshan M, Bora H, Sharma YD (2013) Presence of memory T cells and naturally acquired antibodies in *Plasmodium vivax* malaria-exposed individuals against a group of tryptophan-rich antigens with conserved sequences. *J Infect Dis* 207(1):175–185. <https://doi.org/10.1093/infdis/jis650>
100. Lakshmanan V, Rhee KY, Daily JP (2011) Metabolomics and malaria biology. *Mol Biochem Parasitol* 175(2):104–111. <https://doi.org/10.1016/j.molbiopara.2010.09.008>
101. Weckwerth W, Morgenthal K (2005) Metabolomics: from pattern recognition to biological interpretation. *Drug Discov Today* 10(22):1551–1558. [https://doi.org/10.1016/S1359-6446\(05\)03609-3](https://doi.org/10.1016/S1359-6446(05)03609-3)
102. Ralph SA, van Dooren GG, Waller RF, Crawford MJ, Fraunholz MJ, Foth BJ et al (2004) Tropical infectious diseases: metabolic maps and functions of the *Plasmodium falciparum* apicoplast. *Nat Rev Microbiol* 2(3):203–216. <https://doi.org/10.1038/nrmicro843>
103. Forst CV (2006) Host-pathogen systems biology. *Drug Discov Today* 11(5–6):220–227. [https://doi.org/10.1016/S1359-6446\(05\)03735-9](https://doi.org/10.1016/S1359-6446(05)03735-9)
104. Winzeler EA (2006) Applied systems biology and malaria. *Nat Rev Microbiol* 4(2):145–151. <https://doi.org/10.1038/nrmicro1327>
105. Llinás M, Bozdech Z, Wong ED, Adai AT, DeRisi JL (2006) Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Res* 34 (4):1166–1173. <https://doi.org/10.1093/nar/gkj517>
106. Sherman IW (1979) Biochemistry of *Plasmodium* (malaria parasites). *Microbiol Rev* 43 (4):453–495
107. Vial HJ, Eldin P, Tielens AG, van Hellemond JJ (2003) Phospholipids in parasitic protozoa. *Mol Biochem Parasitol* 126(2):143–154. [https://doi.org/10.1016/s0166-6851\(02\)00281-5](https://doi.org/10.1016/s0166-6851(02)00281-5)
108. Mitamura T, Palacpac NM (2003) Lipid metabolism in *Plasmodium falciparum*-infected erythrocytes: possible new targets for malaria chemotherapy. *Microbes Infect* 5(6):545–552. [https://doi.org/10.1016/s1286-4579\(03\)00070-4](https://doi.org/10.1016/s1286-4579(03)00070-4)
109. Pankova-Kholmyansky I, Flescher E (2006) Potential new antimalarial chemotherapeutics based on sphingolipid metabolism. *Chemotherapy* 52(4):205–209. <https://doi.org/10.1159/000093037>
110. Bansal D, Bhatti HS, Sehgal R (2005) Role of cholesterol in parasitic infections. *Lipids Health Dis* 4:10. <https://doi.org/10.1186/1476-511X-4-10>
111. Planche T, Dzeing A, Ngou-Milama E, Kombila M, Stacpoole PW (2005) Metabolic complications of severe malaria. *Curr Top Microbiol Immunol* 295:105–136. https://doi.org/10.1007/3-540-29088-5_5
112. Lakshmanan V, Rhee KY, Wang W, Yu Y, Khafizov K, Fiser A et al (2012) Metabolomic analysis of patient plasma yields evidence of plant-like alpha-linolenic acid metabolism in *Plasmodium falciparum*. *J Infect Dis* 206(2):238–248. <https://doi.org/10.1093/infdis/jis339>
113. Lian LY, Al-Helal M, Roslani AM, Fisher N, Bray PG, Ward SA et al (2009) Glycerol: an unexpected major metabolite of energy metabolism by the human malaria parasite. *Malar J* 8:38. <https://doi.org/10.1186/1475-2875-8-38>
114. Creek DJ, Chua HH, Cobbold SA, Nijagal B, MacRae JI, Dickerman BK et al (2016) Metabolomics-based screening of the malaria box reveals both novel and established mechanisms of action. *Antimicrob Agents Chemother* 60(11):6650–6663. <https://doi.org/10.1128/AAC.01226-16>
115. Allman EL, Painter HJ, Samra J, Carrasquilla M, Llinas M (2016) Metabolomic profiling of the malaria box reveals antimalarial target pathways. *Antimicrob Agents Chemother* 60 (11):6635–6649. <https://doi.org/10.1128/AAC.01224-16>

116. Cobbold SA, Vaughan AM, Lewis IA, Painter HJ, Camargo N, Perlman DH et al (2013) Kinetic flux profiling elucidates two independent acetyl-CoA biosynthetic pathways in *Plasmodium falciparum*. *J Biol Chem* 288(51):36338–36350. <https://doi.org/10.1074/jbc.M113.503557>
117. Salinas JL, Kissinger JC, Jones DP, Galinski MR (2014) Metabolomics in the fight against malaria. *Mem Inst Oswaldo Cruz* 109(5):589–597. <https://doi.org/10.1590/0074-0276140043>
118. Paulson JC, Blixt O, Collins BE (2006) Sweet spots in functional glycomics. *Nat Chem Biol* 2(5):238–248. <https://doi.org/10.1038/nchembio785>
119. Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, Sasisekharan R (2006) Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology* 16(5):82r–90r. <https://doi.org/10.1093/glycob/cwj080>
120. Gowda DC, Gupta P, Davidson EA (1997) Glycosylphosphatidylinositol anchors represent the major carbohydrate modification in proteins of intraerythrocytic stage *Plasmodium falciparum*. *J Biol Chem* 272(10):6428–6439. <https://doi.org/10.1074/jbc.272.10.6428>
121. Smith TK, Gerold P, Crossman A, Paterson MJ, Borissow CN, Brimacombe JS et al (2002) Substrate specificity of the *Plasmodium falciparum* glycosylphosphatidylinositol biosynthetic pathway and inhibition by species-specific suicide substrates. *Biochemistry* 41(41):12395–12406. <https://doi.org/10.1021/bi0203511>
122. Bushkin GG, Ratner DM, Cui J, Banerjee S, Duraisingh MT, Jennings CV et al (2010) Suggestive evidence for Darwinian selection against asparagine-linked glycans of *Plasmodium falciparum* and *Toxoplasma gondii*. *Eukaryot Cell* 9(2):228–241. <https://doi.org/10.1128/ec.00197-09>
123. Samuelson J, Banerjee S, Magnelli P, Cui J, Kelleher DJ, Gilmore R et al (2005) The diversity of dolichol-linked precursors to Asn-linked glycans likely results from secondary loss of sets of glycosyltransferases. *Proc Natl Acad Sci U S A* 102(5):1548–1553. <https://doi.org/10.1073/pnas.0409460102>
124. Macedo CS, Schwarz RT, Todeschini AR, Previato JO, Mendonça-Previato L (2010) Overlooked post-translational modifications of proteins in *Plasmodium falciparum*: N- and O-glycosylation—a review. *Mem Inst Oswaldo Cruz* 105(8):949–956. <https://doi.org/10.1590/s0074-02762010000800001>
125. von Itzstein M, Plebanski M, Cooke BM, Coppel RL (2008) Hot, sweet and sticky: the glyobiology of *Plasmodium falciparum*. *Trends Parasitol* 24(5):210–218. <https://doi.org/10.1016/j.pt.2008.02.007>
126. Lingelbach K, Kirk K, Rogerson S, Langhorne J, Carucci DJ, Waters A (2004) Molecular approaches to malaria. *Mol Microbiol* 54(3):575–587. <https://doi.org/10.1111/j.1365-2958.2004.04362.x>
127. Baum J, Maier AG, Good RT, Simpson KM, Cowman AF (2005) Invasion by *P. falciparum* merozoites suggests a hierarchy of molecular interactions. *PLoS Pathog* 1(4):e37. <https://doi.org/10.1371/journal.ppat.0010037>
128. Chen Q, Schlichtherle M, Wahlgren M (2000) Molecular aspects of severe malaria. *Clin Microbiol Rev* 13(3):439–450. <https://doi.org/10.1128/cmr.13.3.439-450.2000>
129. Rowe JA, Moulds JM, Newbold CI, Miller LH (1997) *P. falciparum* rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. *Nature* 388(6639):292–295. <https://doi.org/10.1038/40888>
130. Barragan A, Fernandez V, Chen Q, von Euler A, Wahlgren M, Spillmann D (2000) The duffy-binding-like domain 1 of *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) is a heparan sulfate ligand that requires 12 mers for binding. *Blood* 95(11):3594–3599
131. Vogt AM, Winter G, Wahlgren M, Spillmann D (2004) Heparan sulphate identified on human erythrocytes: a *Plasmodium falciparum* receptor. *Biochem J* 381(Pt 3):593–597. <https://doi.org/10.1042/bj20040762>
132. Rowe A, Obeiro J, Newbold CI, Marsh K (1995) *Plasmodium falciparum* rosetting is associated with malaria severity in Kenya. *Infect Immun* 63(6):2323–2326. <https://doi.org/10.1128/iai.63.6.2323-2326.1995>

133. Rowe JA, Shafi J, Kai OK, Marsh K, Raza A (2002) Nonimmune IgM, but not IgG binds to the surface of *Plasmodium falciparum*-infected erythrocytes and correlates with rosetting and severe malaria. *Am J Trop Med Hyg* 66(6):692–699. <https://doi.org/10.4269/ajtmh.2002.66.692>
134. Kupferschmid M, Aquino-Gil MO, Shams-Eldin H, Schmidt J, Yamakawa N, Krzewinski F et al (2017) Identification of O-GlcNAcylated proteins in *Plasmodium falciparum*. *Malar J* 16(1):485. <https://doi.org/10.1186/s12936-017-2131-2>
135. Poliakov LM, Sumenkova DV, Kniazev RA, Panin LE (2011) The analysis of interaction between lipoproteins and steroid hormones. *Biomed Khim* 57(3):308–313. <https://doi.org/10.18097/pbmc20115703308>
136. Lasonder E, Treeck M, Alam M, Tobin AB (2012) Insights into the *Plasmodium falciparum* schizont phospho-proteome. *Microbes Infect* 14(10):811–819. <https://doi.org/10.1016/j.micinf.2012.04.008>
137. Pease BN, Huttlin EL, Jedrychowski MP, Talevich E, Harmon J, Dillman T et al (2013) Global analysis of protein expression and phosphorylation of three stages of *Plasmodium falciparum* intraerythrocytic development. *J Proteome Res* 12(9):4028–4045. <https://doi.org/10.1021/pr400394g>
138. Lasonder E, Green JL, Grainger M, Langsley G, Holder AA (2015) Extensive differential protein phosphorylation as intraerythrocytic *Plasmodium falciparum* schizonts develop into extracellular invasive merozoites. *Proteomics* 15(15):2716–2729. <https://doi.org/10.1002/pmic.201400508>
139. Zuccala ES, Satchwell TJ, Angrisano F, Tan YH, Wilson MC, Heesom KJ et al (2016) Quantitative phospho-proteomics reveals the *Plasmodium* merozoite triggers pre-invasion host kinase modification of the red cell cytoskeleton. *Sci Rep* 6:19766. <https://doi.org/10.1038/srep19766>
140. Alam MM, Solyakov L, Bottrill AR, Flueck C, Siddiqui FA, Singh S et al (2015) Phosphoproteomics reveals malaria parasite protein kinase G as a signalling hub regulating egress and invasion. *Nat Commun* 6:7285. <https://doi.org/10.1038/ncomms8285>
141. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E et al (2008) Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 455(7214):757–763. <https://doi.org/10.1038/nature07327>
142. Aurecochea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B et al (2009) PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res* 37(Database issue):D539–D543. <https://doi.org/10.1093/nar/gkn814>
143. Volkman SK, Sabeti PC, DeCaprio D, Neafsey DE, Schaffner SF, Milner DA Jr et al (2007) A genome-wide map of diversity in *Plasmodium falciparum*. *Nat Genet* 39(1):113–119. <https://doi.org/10.1038/ng1930>
144. Jomaa H, Wiesner J, Sanderbrand S, Altincicek B, Weidemeyer C, Hintz M et al (1999) Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* 285(5433):1573–1576. <https://doi.org/10.1126/science.285.5433.1573>
145. Mu J, Myers RA, Jiang H, Liu S, Ricklefs S, Waisberg M et al (2010) *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nat Genet* 42(3):268–271. <https://doi.org/10.1038/ng.528>
146. Rottmann M, McNamara C, Yeung BK, Lee MC, Zou B, Russell B et al (2010) Spiroindolones, a potent compound class for the treatment of malaria. *Science* 329(5996):1175–1180. <https://doi.org/10.1126/science.1193225>
147. Westenberger SJ, McClean CM, Chattopadhyay R, Dharia NV, Carlton JM, Barnwell JW et al (2010) A systems-based analysis of *Plasmodium vivax* lifecycle transcription from human to mosquito. *PLoS Negl Trop Dis* 4(4):e653. <https://doi.org/10.1371/journal.pntd.0000653>
148. Harbut MB, Patel BA, Yeung BK, McNamara CW, Bright AT, Ballard J et al (2012) Targeting the ERAD pathway via inhibition of signal peptide peptidase for antiparasitic therapeutic design. *Proc Natl Acad Sci U S A* 109(52):21486–21491. <https://doi.org/10.1073/pnas.1216016110>

149. Malmquist NA, Moss TA, Mecheri S, Scherf A, Fuchter MJ (2012) Small-molecule histone methyltransferase inhibitors display rapid antimalarial activity against all blood stage forms in *Plasmodium falciparum*. *Proc Natl Acad Sci U S A* 109(41):16708–16713. <https://doi.org/10.1073/pnas.1205414109>
150. Ganesan K, Ponmee N, Jiang L, Fowble JW, White J, Kamchonwongpaisan S et al (2008) A genetically hard-wired metabolic transcriptome in *Plasmodium falciparum* fails to mount protective responses to lethal antifolates. *PLoS Pathog* 4(11):e1000214. <https://doi.org/10.1371/journal.ppat.1000214>
151. Natalang O, Bischoff E, Deplaine G, Proux C, Dillies MA, Sismeiro O et al (2008) Dynamic RNA profiling in *Plasmodium falciparum* synchronized blood stages exposed to lethal doses of artesunate. *BMC Genomics* 9:388. <https://doi.org/10.1186/1471-2164-9-388>
152. Briolant S, Almeras L, Belghazi M, Boucomont-Chapeaublanc E, Wurtz N, Fontaine A et al (2010) *Plasmodium falciparum* proteome changes in response to doxycycline treatment. *Malar J* 9:141. <https://doi.org/10.1186/1475-2875-9-141>
153. Prieto JH, Koncarevic S, Park SK, Yates J 3rd, Becker K (2008) Large-scale differential proteome analysis in *Plasmodium falciparum* under drug treatment. *PLoS One* 3(12):e4098. <https://doi.org/10.1371/journal.pone.0004098>
154. Amambua-Ngwa A, Tetteh KK, Manske M, Gomez-Escobar N, Stewart LB, Deerrhake ME et al (2012) Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genet* 8(11):e1002992. <https://doi.org/10.1371/journal.pgen.1002992>
155. Mu J, Awadalla P, Duan J, McGee KM, Keebler J, Seydel K et al (2007) Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat Genet* 39(1):126–130. <https://doi.org/10.1038/ng1924>
156. Weedall GD, Conway DJ (2010) Detecting signatures of balancing selection to identify targets of anti-parasite immunity. *Trends Parasitol* 26(7):363–369. <https://doi.org/10.1016/j.pt.2010.04.002>
157. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol* 1(1):E5. <https://doi.org/10.1371/journal.pbio.0000005>
158. Ngwa CJ, Scheuermayer M, Mair GR, Kern S, Brügl T, Wirth CC et al (2013) Changes in the transcriptome of the malaria parasite *Plasmodium falciparum* during the initial phase of transmission from the human to the mosquito. *BMC Genomics* 14:256. <https://doi.org/10.1186/1471-2164-14-256>
159. Tarun AS, Peng X, Dumpit RF, Ogata Y, Silva-Rivera H, Camargo N et al (2008) A combined transcriptome and proteome survey of malaria parasite liver stages. *Proc Natl Acad Sci U S A* 105(1):305–310. <https://doi.org/10.1073/pnas.0710780104>
160. Eksi S, Morahan BJ, Haile Y, Furuya T, Jiang H, Ali O et al (2012) *Plasmodium falciparum* gametocyte development 1 (*Pfgdv1*) and gametocytogenesis early gene identification and commitment to sexual development. *PLoS Pathog* 8(10):e1002964. <https://doi.org/10.1371/journal.ppat.1002964>
161. Hoeijmakers WA, Bartfai R, Stunnenberg HG (2013) Transcriptome analysis using RNA-Seq. *Methods Mol Biol* 923:221–239. https://doi.org/10.1007/978-1-62703-026-7_15
162. Pons N, Chung DW, Le Roch KG (2012) Strand-specific RNA-seq applied to malaria samples. *Methods Mol Biol* 883:59–73. https://doi.org/10.1007/978-1-61779-839-9_4
163. Sorber K, Dimon MT, DeRisi JL (2011) RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Res* 39(9):3820–3835. <https://doi.org/10.1093/nar/gkq1223>
164. Chou ES, Abidi SZ, Teye M, Leliwa-Sytek A, Rask TS, Cobbold SA et al (2018) A high parasite density environment induces transcriptional changes and cell death in *Plasmodium falciparum* blood stages. *FEBS J* 285(5):848–870. <https://doi.org/10.1111/febs.14370>

165. Kumar K, Srinivasan P, Nold MJ, Moch JK, Reiter K, Sturdevant D et al (2017) Profiling invasive *Plasmodium falciparum* merozoites using an integrated omics approach. *Sci Rep* 7 (1):17146. <https://doi.org/10.1038/s41598-017-17505-9>
166. Balaji S, Babu MM, Iyer LM, Aravind L (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res* 33(13):3994–4006. <https://doi.org/10.1093/nar/gki709>
167. Bancells C, Llorà-Batlle O, Poran A, Nötzel C, Rovira-Graells N, Elemento O et al (2019) Revisiting the initial steps of sexual development in the malaria parasite *Plasmodium falciparum*. *Nat Microbiol* 4(1):144–154. <https://doi.org/10.1038/s41564-018-0291-7>
168. Vontas J, Siden-Kiamos I, Papagiannakis G, Karras M, Waters AP, Louis C (2005) Gene expression in *Plasmodium berghei* ookinetes and early oocysts in a co-culture system with mosquito cells. *Mol Biochem Parasitol* 139(1):1–13. <https://doi.org/10.1016/j.molbiopara.2004.03.003>
169. Bozdech Z, Mok S, Hu G, Imwong M, Jaidee A, Russell B et al (2008) The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites. *Proc Natl Acad Sci U S A* 105(42):16290–16295. <https://doi.org/10.1073/pnas.0807404105>
170. Pattaradilokrat S, Cheesman SJ, Carter R (2008) Congenicity and genetic polymorphism in cloned lines derived from a single isolate of a rodent malaria parasite. *Mol Biochem Parasitol* 157(2):244–247. <https://doi.org/10.1016/j.molbiopara.2007.10.011>
171. Dharia NV, Bright AT, Westenberger SJ, Barnes SW, Batalov S, Kuhlen K et al (2010) Whole-genome sequencing and microarray analysis of ex vivo *Plasmodium vivax* reveal selective pressure on putative drug resistance genes. *Proc Natl Acad Sci U S A* 107(46):20045–20050. <https://doi.org/10.1073/pnas.1003776107>
172. Roobsoong W, Roytrakul S, Sattabongkot J, Li J, Udomsangpetch R, Cui L (2011) Determination of the *Plasmodium vivax* schizont stage proteome. *J Proteome* 74(9):1701–1710. <https://doi.org/10.1016/j.jprot.2011.03.035>
173. Chan ER, Menard D, David PH, Ratsimbaoa A, Kim S, Chim P et al (2012) Whole genome sequencing of field isolates provides robust characterization of genetic diversity in *Plasmodium vivax*. *PLoS Negl Trop Dis* 6(9):e1811. <https://doi.org/10.1371/journal.pntd.0001811>
174. Auburn S, Marfurt J, Maslen G, Campino S, Ruano Rubio V, Manske M et al (2013) Effective preparation of *Plasmodium vivax* field isolates for high-throughput whole genome sequencing. *PLoS One* 8(1):e53160. <https://doi.org/10.1371/journal.pone.0053160>
175. Winter DJ, Pacheco MA, Vallejo AF, Schwartz RS, Arevalo-Herrera M, Herrera S et al (2015) Whole genome sequencing of field isolates reveals extensive genetic diversity in *Plasmodium vivax* from Colombia. *PLoS Negl Trop Dis* 9(12):e0004252. <https://doi.org/10.1371/journal.pntd.0004252>
176. Auburn S, Bohme U, Steinbiss S, Trimarsanto H, Hostetler J, Sanders M et al (2016) A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of *pir* genes. *Wellcome Open Res* 1:4. <https://doi.org/10.12688/wellcomeopenres.9876.1>
177. Zhu L, Mok S, Imwong M, Jaidee A, Russell B, Nosten F et al (2016) New insights into the *Plasmodium vivax* transcriptome using RNA-Seq. *Sci Rep* 6:20498. <https://doi.org/10.1038/srep20498>
178. Hester J, Chan ER, Menard D, Mercereau-Puijalon O, Barnwell J, Zimmerman PA et al (2013) De novo assembly of a field isolate genome reveals novel *Plasmodium vivax* erythrocyte invasion genes. *PLoS Negl Trop Dis* 7(12):e2569. <https://doi.org/10.1371/journal.pntd.0002569>



Amrendra Nath Pathak, Lalit Kumar Singh, and Esha Dwivedi

Abstract

The microorganisms adapt themselves to the changing environments. A sequence of complex, when a pathogen enters its host, intracellular associations shape the infection's outcome. It is important to understand the host-pathogen relationships in order to evolve the interventions. Proteomics since the last many years have been key contributors to exploration, and it helps to apprehend the anti-pathogen relationships between hosts and pathogens. After the extension of the comprehension, we address the strategies of proteome organization after an outbreak. It helps in characterizing the regulation of the interaction between hosts and organisms via shifts in the concentration of the proteins and the posttranslational alterations.

Keywords

Metabolomics · Cell organ · Posttranslational modifications · Illness · Microorganism

A. N. Pathak (✉)

Centre of Research for Development, Parul University, Vadodra, Gujrat, India

e-mail: amrendra.pathak@paruluniversity.ac.in

L. K. Singh · E. Dwivedi

Department of Biochemical Engineering, School of Chemical Technology, Harcourt Butler

Technical University Kanpur (UP), Kanpur, UP, India

e-mail: lkumar@hbtu.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

507

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_27

27.1 Proteomics: The Consequential Method to Contemplate Contagious Disease Disorder

One of the most provocative and engrossing connections between the interactions of hosts and pathogens is the tantalizing facets of our life. These communications have been formed since the past years, and the hosts are evolving the defence mechanisms in antithetical to the pathogens. The various process has been required to prevail the hosts. Also benefitting via pathogens, many pathogens are effectual and pivotal. The agencies with a multitudinous of human illness recognize the relationships between hosts and the pathogens for the prevention and care from illness. The associations of the host and the microorganism hold the position frequently at the pace of the molecular level. This subjects the analysis of various bacteria, viruses, and intracellular toxins. The essential part of the pathogens is to counteract the host with the cycle of the replication. A large number of years of studies have been done to interpretate the microbial life cycle. The various problems are very much complex in order to discuss the application of classical molecular biology approaches. The rise of the associated diseases involves the identification of the hospital and the growth of the analysis [1]. The growth of the viruses and the bacteria that are resistant to the drugs is needed to explore the routes that may be aimed to block the transmission of the pathogens. The looming and impending viruses do not persist with effective medications [2]. The arrangement of the reproduction differs, and these pathogens perform many functions such as the joining of the cell and distribution to adjacent cells (Fig. 27.1).

The various steps for the cytosolic replication viruses which are non-enveloped are shown in the above-mentioned figure. The entry of the virus takes place by the endocytosis. General advances have appeared for cytosolic-reproducing non-wrapped infections, atomic imitating encompassed infections, and cytosolic microscopic organisms. Infection joins and makes its entrance into the cavity via phagocytosis, pinocytosis, caveolae, or with the association of a cell in which the vector can be propagated. Viral qualities are then communicated to deliver viral proteins. Viral proteins encourage safe avoidance, viral genome replication, viral genome encapsidation, and envelopment. Completely amassed irresistible infections are emitted via destruction or dissolution of the cells and also via cell transport, cellular secretion, and active transport. At that point, since microorganisms get supplements legitimately from the host cytosol, the microbes recreate inside the liquid found inside the cells. Second, the rise of medication safe infections and microorganisms features the need to find the pathways that can be directed to obstruct the spread of microorganisms. Finally, therapeutically undermining infections that have been researched for a long time still persevere with no reasonable medicines or immunizations [3]. The incorporation of proteomics with other biochemical and atomic science strategies has extended the collection of instruments to consider microbe contaminations. The inheritable gene of the chromosome contemplates give symmetrical data that supplements proteomic examinations to accomplish a frameworks level comprehension of the contamination measure. The innovations yield extra knowledge into the mechanics of host-microbe connections.

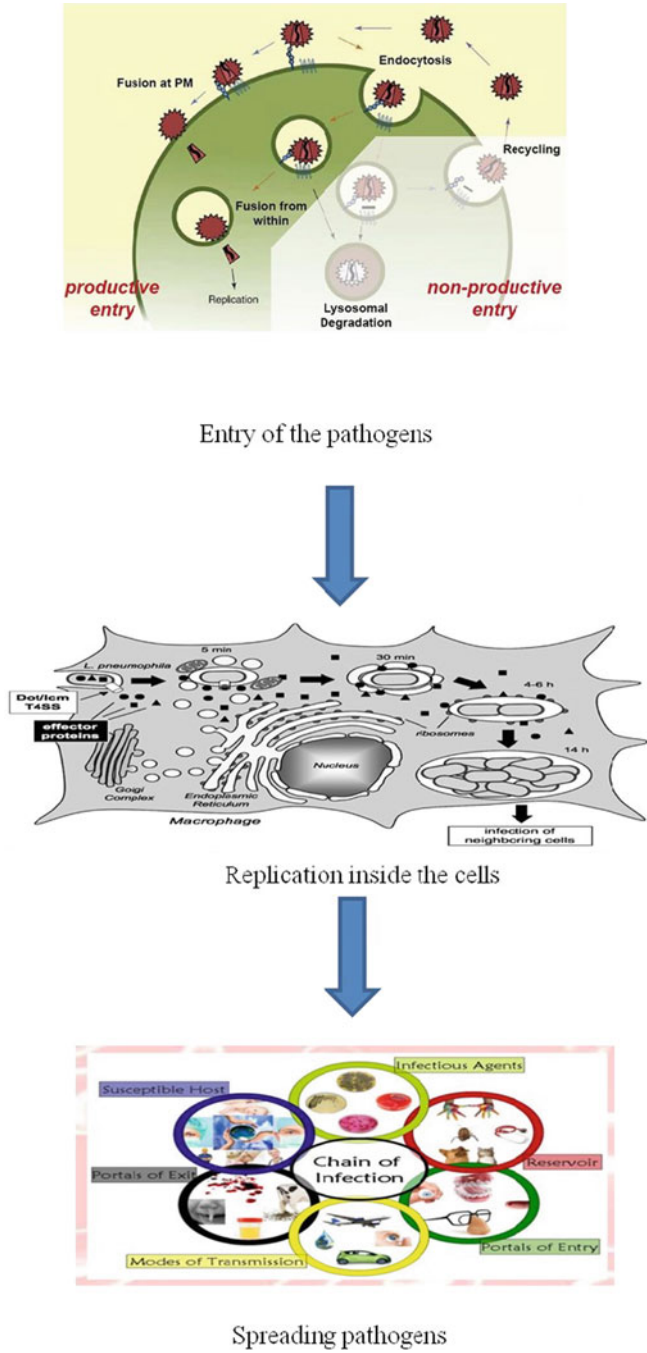


Fig. 27.1 Proteomic methods used to analyse pathogenic pathogens

We additionally talk about bioinformatics apparatuses that structure a vital some portion of proteomics to increment the intensity of revelation and capacity to decipher enormous datasets.

27.2 Host-Microorganism Protein-Protein Connections

Distinguishing proof of these protein-protein associations (PPIs) isn't just basic for the comprehension of the science of disease but however can likewise highlight novel targets in medicines against human microorganisms [4]. In IP-MS, using either a neutralizer elevated besides the substances and processes that originate from within a system such as an organism, tissue, or cell or antigenic determinant labelling of the large number of the polymers of the amino acids of intrigue, a protein of intrigue is disconnected using a neutralizer against the epitope. MS separates the protein of intrigue and co-disconnected cooperating proteins at that point. IP-MS focuses on studies in relevant cell model frameworks and with respect to viral disease when analysing host-microorganism affiliations. IP-MS course and the application can decide modifications by taking into account the connections of the various polymers of the amino acids for the time being taken into account with the contamination to portray conceivable fluctuation, inconsistency in the host protein capacities. At first showed for examining the RNA infection sindbis [5], this methodology was later applied to different infections, for example, the RNA infection respiratory syncytial infection [6] and the DNA infections human cytomegalovirus (HCMV) and pseudorabies infection (PRV) [7]. Also, from the host point of view, IP-MS has assisted with characterizing instruments of cell protection [8] and to recognize protein area subordinate communications and capacities for having antiviral components [9]. Despite the fact that IP-MS has been effectively utilized to consider a few infections, there are still difficulties related with this strategy. A few investigations have used ectopic articulation of labelled viral proteins outside the setting of contamination to get data of potential viral-have PPIs that can be sought after with organic examinations. For instance, this methodology was indicated significant for considering the capacity of the Ebola infection network protein, VP40 [10]. Another example is the interactome of each of the 18 human immunodeficiency infection (HIV) proteins, which have anticipated very nearly 500 microbe communications [11]. Regarding measuring the distinguished host-microbe cooperation, most IP-MS contemplates have depended on mark free MS measurement (e.g. ghastly checking), which is basic, is adaptable, and can be applied to any natural framework. Be that as it may, marking MS methodologies give more precise evaluation of PPI information and can be utilized to look at uninfected and tainted examples in a similar MS try (Fig. 27.2).

The naming can happen at the protein level, through the utilization of stable isotope naming of amino acids in cell culture (SILAC), or at the peptide level, through fuse of pair mass labels (TMT) or other isobaric labels [12]. In contemplating host-infection associations, SILAC was utilized to control for bogus positive PPI IDs, for example, when examining hepatitis C infection [13]. Mark free

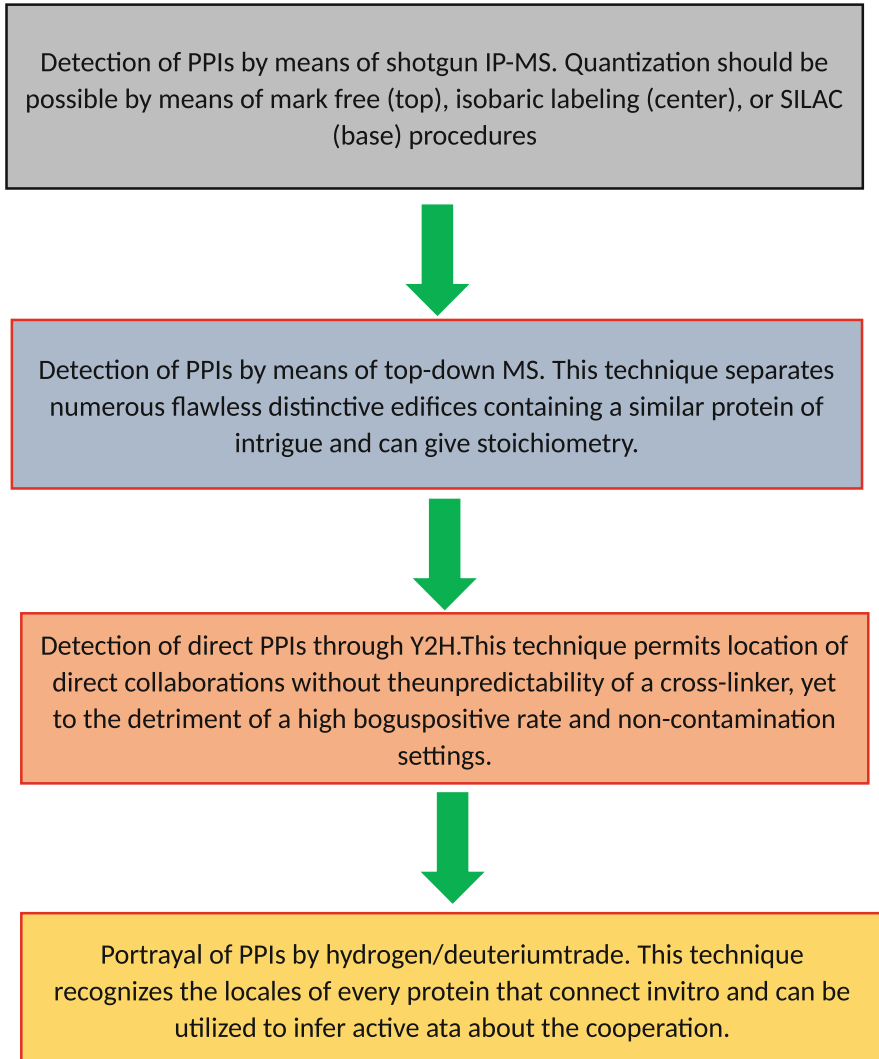


Fig. 27.2 Proteomic instruments to examine protein-protein connections in pathogenic contaminations

and isotopic naming examinations are not commonly restrictive, and a few investigations have joined SILAC with mark free IP-MS to extraordinary impact [14]. For instance, a consolidated examination was utilized to decide both explicit cooperation of histone deacetylases by mark free strategies and the general steadiness of these collaborations by SILAC [15]. Such methodologies can hence be extended to give significant data about unique host-microbe connections. As demonstrated over, one restriction to IP-MS datasets is the presence of vaguely

collaborating proteins that co-sanitize with the protein of intrigue. One of the most significant changes seen in the microbe the contaminations that can trigger noteworthy changes in protein plenitudes inside a cell, and the foundation of vague affiliations can be very not the same as the one saw in a uninfected cell. A few accessible PC calculations exist that utilize information from control and test segregations to help channel bogus positive PPIs [16]. One such calculation is the criticalness examination of interactome [17], which allocates communication explicitness scores to channel low-certainty connections. Informatics draws near can likewise be utilized to additionally refine distinguished communications, for instance, by giving extra controls to vague affiliations, for example, the impurity storehouse for liking refinement [18]. Regular assets for network representation incorporate STRING [19] and Cytoscape [20], and we direct the per users toward a convention managing clients through IP-MS information investigation [21].

27.3 Examination of Unblemished Protein Buildings

So as to complete various capacities, proteins oftentimes exist at the same time inside unmistakable protein edifices. In this way, in spite of the fact that IP-MS offers inventories of protein collaborations, it midpoints together different protein buildings that contain a similar protein of intrigue. Until this point in time, in the setting of irresistible ailment, this procedure has been applied principally to singular microorganism proteins, for example, the hepatitis C infection pore protein p7 [22], and pathogenic edifices reconstituted in vitro, for example, the Norwalk infection-like particles [23]. Notwithstanding, top-down MS has not been applied to examining host-microorganism edifices. The capacity to investigate high subatomic mass edifices stays for testing; however MS instrumentation upgrades are consistently broadening the mass range in which these investigations can be applied.

27.4 Distinguishing Direct Collaborations

While the strategies talked about above give fair-minded recognition of collaborations (IP-MS) and data about the unpredictable stoichiometry (top-down MS), these methodologies can't arrange PPIs as immediate or aberrant. Y2H was additionally used to clarify direct PPIs between EHEC and the human host cells [24]. A drawback for Y2H is its generally high bogus positive rate due to the non-physiological articulation of proteins in cell compartments in which they may not regularly be communicated. Moreover, since microorganism proteins are communicated outside the setting of a disease, numerous conceivably pertinent communications can be missed. The ongoing report exploited these cross-connecting apparatuses and computational turn of events (i.e. XLinkDB) to create a huge dataset of direct cooperation between human lung cells and *Acinetobacter baumannii*, a subset of which were demonstrated to be significant in bacterial attack [25]. The utilization of cross-linkers isn't restricted to recognizing PPIs during

contamination. Photograph cross-connecting was utilized to catch RNA-protein communications, giving both stoichiometric and basic data about the start of HIV viral genome bundling [26]. A few examinations utilized cross-connecting MS to recognize proteins that quandy to viral RNA during polio [27] furthermore, dengue [28] infection contaminations.

27.5 Transient Investigation of the Tainted Cell Proteome

Considering worldly proteome modifications has gotten a mainstream approach because of the accessibility of settled conventions and present-day MS instrumentation. When the overall quantitative values are gathered, the hugeness and size of the differential protein wealth are surveyed with measurable strategies, for example, t-test, examination of difference (ANOVA), or more advanced straight models accessible in programming bundles [17]. Extra bioinformatics examinations are needed to connect the proteins with explicit organic pathways and cell capacities. Extra bioinformatics investigations are needed to correspond the proteins with explicit organic pathways and cell capacities. Normally, various levelled grouping is utilized to distinguish sets of proteins with comparative worldly profiles upon contamination. These bunches are then dependent upon utilitarian examination by an assortment of bioinformatics instruments, including quality cosmology investigation, pathway investigation [29], network examination [20], or a blend of these [30]. Transient proteome investigations have been effective in distinguishing pathways directed by the microorganism and key proteins associated with pathogenicity. For instance, infections rely upon cell digestion and furthermore, have obtained systems to direct it for energy creation furthermore, lipid union, among different cycles. Since contaminations actuate wide proteome modifications, contemplates were additionally planned with a limited spotlight on individual pathogenic proteins [31]. So far essentially utilized for cell culture frameworks, worldly proteomic examinations during disease have been effectively applied for in vivo examinations in creature models tested with infections and microbes [32]. An ongoing report exploited these cross-connecting instruments and computational turn of events (i.e. XLinkDB) to create a huge dataset of direct connections between human lung cells and *Acinetobacter baumannii*, a subset of which were demonstrated to be significant in bacterial attack [25]. The utilization of cross-linkers isn't restricted to recognizing PPIs during contamination. Photograph cross-connecting was utilized to catch RNA-protein connections, giving both stoichiometric and basic data about the start of HIV viral genome bundling [26]. The utilization of cross-linkers isn't restricted to recognizing PPIs during communications, giving both stoichiometric and auxiliary data about the start of HIV viral genome bundling [33]. A few investigations utilized cross-connecting MS to distinguish proteins that quandy to viral RNA during polio [27] furthermore, dengue [28] infection diseases. The host proteins distinguished in each investigation were exceptional to the individual viral disease, and resulting knockdown investigations exhibited the need of these proteins for productive viral cycles. The

assessment of RNA-protein connections by MS vows to further extend our comprehension of post-transcriptional guideline measures that may assume significant functions during pathogenic contamination.

27.6 Microorganism Incited Proteome Adjustments in Reality

The creation, corruption, and spatial redesign of proteins are integral for the replication of microorganism. The host additionally reacts to the microbe attack through worldwide adjustments in the proteome association, significant for mounting successful safeguards. For instance, intrinsic resistant and stress reactions to pathogenic intrusion can trigger guideline of tens-to-many proteins [34].

27.7 Wordly Examination of the Contaminated Cell Proteome

Considering worldly proteome modifications has gotten a mainstream approach because of the accessibility of settled conventions and present-day MS instrumentation. Extra bioinformatics examinations are needed to relate the proteins with explicit organic pathways and cell capacities. Ordinarily, various levelled groupings are utilized to distinguish sets of proteins with comparable worldly profiles upon contamination. Fleeting proteome investigations have been effective in recognizing pathways managed by the microorganism and key proteins associated with pathogenicity. For instance, infections rely upon cell digestion also and have obtained instruments to manage it for energy creation also, lipid union, among different cycles. Expansive changes in proteins engaged with digestion guideline have been accounted for from fleeting proteomic investigations of human-significant infections, for example, the as of late reappeared chikungunya infection [35], HCMV [36], flaviviruses [37], and HCV [38]. Post-translational adjustments (PTMs) alter protein functions through ameliorations in protein interactions, stability, activity, and subcellular localization [39].

27.8 Different Varieties of Post-translational Adjustments That Are Applicable in Terms of the Infection

The different types of alteration and the variation changes are applicable in contexture and relation relating to contagion and the disease. The modifications, alterations, changes, and the modifications are applicable at some stage in a number of tiers of the pathogen lifestyle cycle (Fig. 27.3).

Throughout the entrance and the passage, the lipid layer surrounding viruses combines with the cell membrane through fervid and aggressive lectins, mucins, and proteins on the outermost layer of the many types of the viruses. The large-scale substantial reaction in which a carbohydrate, i.e. a glycosyl donor, is attached to a hydroxyl group is antediluvianated and is determined and remarked in envelope

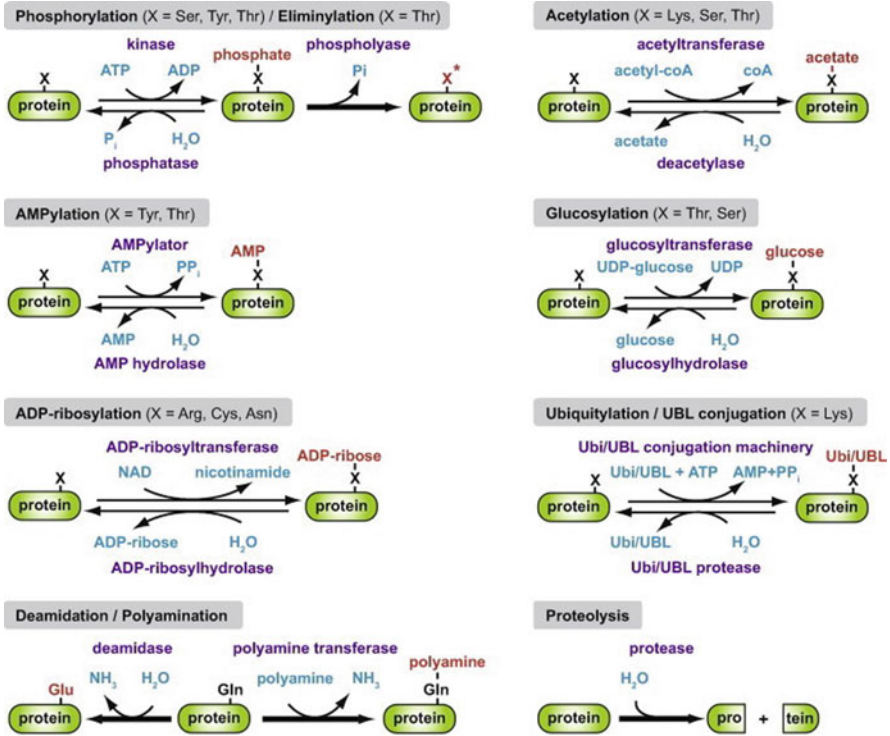


Fig. 27.3 Post-translational adjustments worried in the context of the contamination

biomolecules referring to cold sores, counting HSV-1, HCMV, varicella zoster virus (VZV), and Epstein-Barr virus (EBV). These are manifested and seem to be imperative considering its pathogens ability to infect or damage the host [40]. The various omics approaches used to examine the host and the pathogen relationship are mentioned in Fig. 27.4:

27.9 Shotgun Metagenomics

Metagenomic data canalize the discovery of genetic pathways with unique roles associated and should also provide practical knowledge on the future technical capacities of communities of microbials using antifungal metagenomics, activities, search ability, and extracellular secondary production.

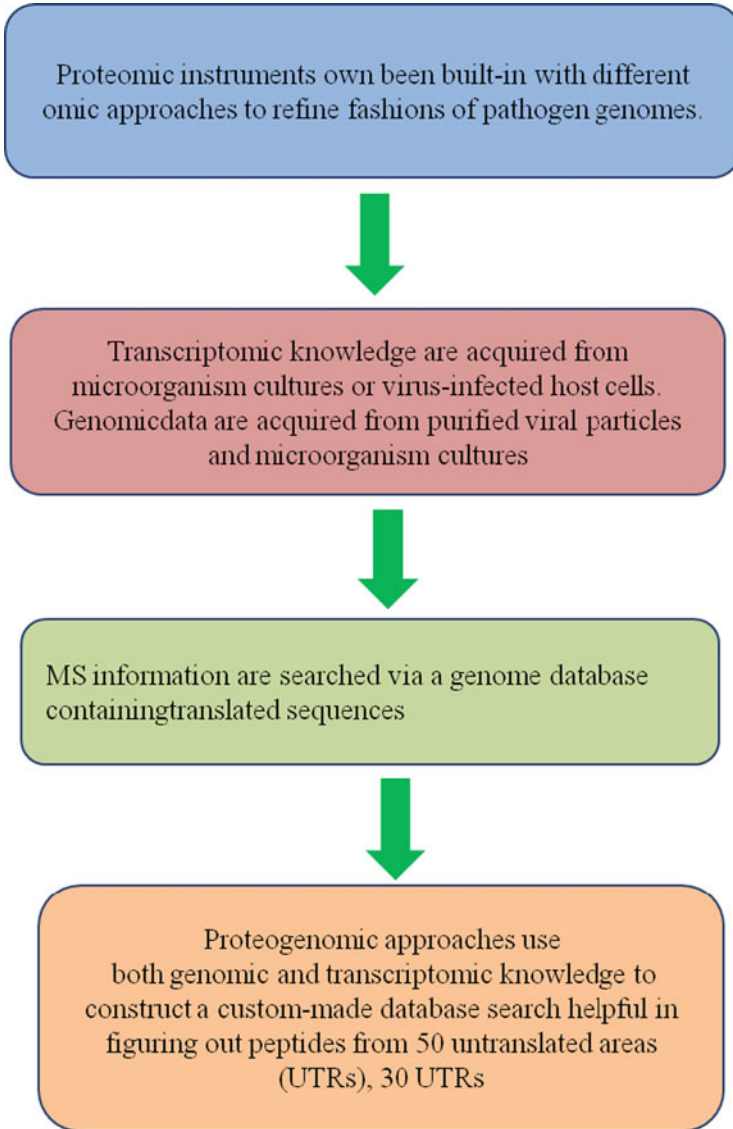


Fig. 27.4 The various omics approaches used to examine the host and the pathogen relationship

27.10 Metatranscriptomics

Metatranscriptomics is the analysis and the community profiles are meaningless for the discovery of genes or regulated genetic pathways that are up or downresponse to this method and canal so unravel functional, and a pathogen infection such as the

responses involved in bacterial-host interactions. Additional traits are consistent with the expression of adhesion genes or eukaryotes of bacterial colonization and attachment. Experimental metatranscriptomics may be a good approaching method and the laboratory environments, exposed to the amphibians in Bd.

27.11 Important Considerations and Directions for Future

In addition to probiotic treatment, here is the introduction of the species still in captivity, one big challenge. The research of human infectious disease has been ongoing in recent years. The contribution of proteomics has greatly gained from proteomic approaches focused on quantitative MS (e.g. TMT labelling and SILAC) for sensitive people and has been well known. The detection of PPIs between host and host pathways is the pathogen proteins and their complex regulation and infection course. Future studies of such PPIs will make use of the current platforms for proteomics, and while still taking advantage of the constantly enhancing quantitative approaches to MS, particularly for the study of interactomes, and also in the incorporation of spatial and temporal resolution. The continued development of analysis and interpretation algorithms, protein abundance data, interactions, and PTMs will facilitate analysis of the pathogenic virus underlying biology. Technologies offer resources for researchers to receive a holistic image of the relationship between the host and the pathogen with the objective of obtaining a clearer knowledge of disease processes for the future and therapeutic aim exploration.

References

1. Morens DM, Fauci AS (2012) Emerging infectious diseases in 2012: 20 years after the institute of medicine report. *MBio* 3:e00494–e00412
2. Su MA, Huang YT, Chen IT, Lee DY, Hsieh YC, Li CY, Ng TH, Liang SY, Lin SY, Huang SW, Chiang YA, Yu HT, Khoo KH, Chang GD, Lo CF, Wang HC (2014) An invertebrate Warburg effect: a shrimp virus achieves successful replication by altering the host metabolome via the PI3K-Akt-mTOR pathway. *PLoS Pathog* 10:e1004196
3. Rieder F, Steininger C (2014) Cytomegalovirus vaccine: phase II clinical trial results. *Clin Microbiol Infect* 20(Suppl. 5):95–102
4. Lum KK, Cristea IM (2016) Proteomic approaches to uncovering virus-host protein interactions during the progression of viral infection. *Expert Rev Proteomics* 13:325–340
5. Cristea IM, Carroll JW, Rout MP, Rice CM, Chait BT, MacDonald MR (2006) Tracking and elucidating alphavirus-host protein interactions. *J Biol Chem* 281:30269–30278
6. Wu W, Tran KC, Teng MN, Heesom KJ, Matthews DA, Barr JN, Hiscox JA (2012) The interactome of the human respiratory syncytial virus NS1 protein highlights multiple effects on host cell biology. *J Virol* 86:7777–7789
7. Kramer T, Greco TM, Taylor MP, Ambrosini AE, Cristea IM, Enquist LW (2012) Kinesin-3 mediates axonal sorting and directional transport of alpha herpes virus particles in neurons. *Cell Host Microbe* 12:806–814

8. Diner BA, Lum KK, Javitt A, Cristea IM (2015) Interactions of the antiviral factor interferon gamma-inducible protein 16 (IFI16) mediate immune signaling and herpes simplex virus-1 immunosuppression. *Mol Cell Proteomics* 14:2341–2356
9. Diner BA, Lum KK, Toettcher JE, Cristea IM (2016) Viral DNA sensors IFI16 and cyclic GMP-AMP synthase possess distinct functions in regulating viral gene expression, immune defenses, and apoptotic responses during herpes virus infection. *MBio* 7:e01553–e01516
10. Yamayoshi S, Noda T, Ebihara H, Goto H, Morikawa Y, Lukashevich IS, Neumann G, Feldmann H, Kawaoka Y (2008) Ebola virus matrix protein VP40 uses the COPII transport system for its intracellular transport. *Cell Host Microbe* 3:168–177
11. Jager S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, Shales M, Mercenne G, Pache L, Li K, Hernandez H, Jang GM, Roth SL, Akiva E, Marlett J, Stephens M, D’Orso I, Fernandes J, Fahey M, Mahon C et al (2012) Global landscape of HIV-human protein complexes. *Nature* 481:365–370
12. Bantscheff M, Lemeer S, Savitski MM, Kuster B (2012) Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* 404:939–965
13. Gerold G, Meissner F, Bruening J, Welsch K, Perin PM, Baumert TF, Vondran FW, Kaderali L, Marcotrigiano J, Khan AG, Mann M, Rice CM, Pietschmann T (2015) Quantitative proteomics identifies serum response factor binding protein 1 as a host factor for hepatitis C virus entry. *Cell Rep* 12:864–878
14. Auweter SD, Bhavsar AP, de Hoog CL, Li Y, Chan YA, van der Heijden J, Lowden MJ, Coombes BK, Rogers LD, Stoykov N, Foster LJ, Finlay BB (2011) Quantitative mass spectrometry catalogues *Salmonella* pathogenicity island-2 effectors and identifies their cognate host binding partners. *J Biol Chem* 286:24023–24035
15. Joshi P, Greco TM, Guise AJ, Luo Y, Yu F, Nesvizhskii AI, Cristea IM (2013) The functional interactome landscape of the human histone deacetylase family. *Mol Syst Biol* 9:672
16. Armean IM, Lilley KS, Trotter MW (2013) Popular computational methods to assess multiprotein complexes derived from label-free affinity purification and mass spectrometry (AP-MS) experiments. *Mol Cell Proteomics* 12:1–13
17. Choi H, Larsen B, Lin ZY, Breitkreutz A, Mellacheruvu D, Fermin D, Qin ZS, Tyers M, Gingras AC, Nesvizhskii AI (2011) SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat Methods* 8:70–73
18. Mellacheruvu D, Wright Z, Couzens AL, Lambert JP, St-Denis NA, Li T, Miteva YV, Hauri S, Sardi ME, Low TY, Halim VA, Bagshaw RD, Hubner NC, Al-Hakim A, Bouchard A, Faubert D, Fermin D, Dunham WH, Goudreau M, Lin ZY et al (2013) The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods* 10:730–736
19. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43:D447–D452
20. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR et al (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2:2366–2382
21. Morris JH, Knudsen GM, Verschueren E, Johnson JR, Cimermancic P, Greninger AL, Pico AR (2014) Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions. *Nat Protoc* 9:2539–2554
22. Konijnenberg A, Bannwarth L, Yilmaz D, Kocer A, Venien-Bryan C, Sobott F (2015) Top-down mass spectrometry of intact membrane protein complexes reveals oligomeric state and sequence information in a single experiment. *Protein Sci* 24:1292–1300
23. Shoemaker GK, van Duijn E, Crawford SE, Utrecht C, Baclayon M, Roos WH, Wuite GJ, Estes MK, Prasad BV, Heck AJ (2010) Norwalk virus assembly and stability monitored by mass spectrometry. *Mol Cell Proteomics* 9:1742–1751

24. Blasche S, Arens S, Ceol A, Sisler G, Schmidt MA, Hauser R, Schwarz F, Wuchty S, Aloy P, Uetz P, Stradal T, Koegl M (2014) The EHEC-host interactome reveals novel targets for the translocated intimin receptor. *Sci Rep* 4:7531
25. Schweppe DK, Harding C, Chavez JD, Wu X, Ramage E, Singh PK, Manoil C, Bruce JE (2015) Host-microbe protein interactions during bacterial infection. *Chem Biol* 22:1521–1530
26. Kenyon JC, Prestwood LJ, Lever AM (2015) A novel combined RNA-protein interaction analysis distinguishes HIV-1 gag protein binding sites from structural change in the viral RNA leader. *Sci Rep* 5:14369
27. Lenarcic EM, Landry DM, Greco TM, Cristea IM, Thompson SR (2013) Thiouracil cross-linking mass spectrometry: a cell-based method to identify host factors involved in viral amplification. *J Virol* 87:8697–8712
28. Viktorovskaya OV, Greco TM, Cristea IM, Thompson SR (2016) Identification of RNA binding proteins associated with dengue virus RNA in infected cells reveals temporally distinct host factor requirements. *PLoS Negl Trop Dis* 10:e0004921
29. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44:D457–D462
30. Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* 8:1551–1566
31. Wood JJ, Boyne JR, Paulus C, Jackson BR, Nevels MM, Whitehouse A, Hughes DJ (2016) ARID3B: a novel regulator of the Kaposi's sarcoma-associated herpes virus lytic cycle. *J Virol* 90:9543–9555
32. Lopez V, Villar M, Queiros J, Vicente J, Mateos-Hernandez L, Diez-Delgado I, Contreras M, Alves PC, Alberdi P, Gortazar C, de la Fuente J (2016) Comparative proteomics identifies host immune system proteins affected by infection with *Mycobacterium bovis*. *PLoS Negl Trop Dis* 10:e0004541
33. Khatri K, Klein JA, White MR, Grant OC, Leymarie N, Woods RJ, Hartshorn KL, Zaia J (2016) Integrated omics and computational glycobiology reveal structural basis for influenza A virus glycan microheterogeneity and host interactions. *Mol Cell Proteomics* 15:1895–1912
34. Janssens S, Pulendran B, Lambrecht BN (2014) Emerging functions of the unfolded protein response in immunity. *Nat Immunol* 15:910–919
35. Abere B, Wikan N, Ubol S, Auewarakul P, Paemanee A, Kittisenachai S, Roytrakul S, Smith DR (2012) Proteomic analysis of chikungunya virus infected microbial cells. *PLoS One* 7: e34800
36. Weekes MP, Tomasec P, Huttlin EL, Fielding CA, Nusinow D, Stanton RJ, Wang EC, Aicheler R, Murrell I, Wilkinson GW, Lehner PJ, Gygi SP (2014) Quantitative temporal viromics: an approach to investigate host-pathogen interaction. *Cell* 157:1460–1472
37. Pastorino B, Boucomont-Chapeaublanc E, Peyrefitte CN, Belghazi M, Fusai T, Rogier C, Tolou HJ, Almeras L (2009) Identification of cellular proteome modifications in response to West Nile virus infection. *Mol Cell Proteomics* 8:1623–1637
38. Diamond DL, Syder AJ, Jacobs JM, Sorensen CM, Walters KA, Proll SC, McDermott JE, Gritsenko MA, Zhang Q, Zhao R, Metz TO, Camp DG II, Waters KM, Smith RD, Rice CM, Katze MG (2010) Temporal proteome and lipidome profiles reveal hepatitis C virus-associated reprogramming of hepatocellular metabolism and bioenergetics. *PLoS Pathog* 6:e1000719
39. Ribet D, Cossart P (2010) Pathogen-mediated posttranslational modifications: a re-emerging field. *Cell* 143:694–702
40. Bagdonaite I, Norden R, Joshi HJ, King SL, Vakhrushev SY, Olofsson S, Wandall HH (2016) Global mapping of O-glycosylation of varicella zoster virus, human cytomegalovirus, and Epstein-Barr virus. *J Biol Chem* 291:12014–12028



Pathogen-Omics: Challenges and Prospects in Research and Clinical Settings **28**

Dyuti Purkait, Saif Hameed, and Zeeshan Fatima

Abstract

“Omics” can be described as a shorthand term for several recent technologies which are in totality for molecular composition of living organism. Generally “omics” is used as suffix to any terminology in the biological sciences which refers to comprehensive, or global, study of biological molecules. The human infectious diseases present ever pressing challenge to the scientific community. The advancements in omics technology present a potential alternative to combat the infectious diseases and comprehend human pathobiology. Recent omics-based approaches have unveiled the molecular mechanisms behind microbial infections and help in elucidating next-generation biomarkers for early diagnosis, prognosis, and targeted therapeutics. In the present chapter, the authors describe the biological impact of the various omics approaches that have been integrated to study infectious diseases and outline many of the targets and processes that can be assessed as part of a comprehensive omics analysis of the pathogens.

Keywords

Human pathogens · Omics · Genomics · Transcriptomics · Proteomics · Lipidomics · Metabolomics

28.1 Introduction

“Omics” can be described as a shorthand term for several recent technologies which are used to examine holistic view of molecular composition of living organism. The term “omics” is used as suffix to a molecular term in the biological sciences and

D. Purkait · S. Hameed · Z. Fatima (✉)

Amity Institute of Biotechnology, Amity University Haryana, Gurugram, Haryana, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*, https://doi.org/10.1007/978-981-16-0691-5_28

refers to comprehensive, or global, study of biological molecules [1]. As the Human Genome Project (HGP) was completed in 2001, the idea evolved that the field of molecular biology study should explore biological molecules from isolates toward a broad analysis of large sets of such molecules [2, 3]. Complex biological processes are regulated not only by DNA sequence as proven by the HGP, which triggered many rapid developments in the field of molecular biology in several aspects which were together known by the term omics. The first omics discipline to appear was genomics which is the field to focus on the study of entire genomes, followed by transcriptomics which focus on RNA, proteomics focusing on protein, and metabolomics which is focused on the metabolome.

The omics field is largely dependent on advance technologies which made possible cost-efficient, high-throughput analysis of biological molecules. Example of which that was developed in the late 1990s is the “expression array,” which is based on hybridization of cDNA to arrays of oligonucleotide with probe [4]. The omics term was coined by Marc Wilkins in 1994. Because of advanced techniques like high-resolution two-dimensional electrophoresis, the study of proteomics is possible, and the advantage of the omics study is to reveal specific results that promote understanding. But with advancement of technology, mass spectrometry-based proteomics and metabolomics studies are capable of comprehensive measurement of proteins and metabolomes which provides valuable insight into the molecular mechanisms and dynamics of biological processes. Keeping view of the immense potential of the omics technology, they have been extensively used in various branches of medical and health science. Omics technology is used in the process of screening, diagnosis, and prognosis to understand the etiology of disease and also useful for developing biomarker [5]. In addition, omics technology has great usefulness in drug discovery and toxicity assessment. Each type of omics data can be useful for identifying markers of the disease process as well as to get an insight of comparison data of biological pathways between the disease and control individuals of the study. Integration of different omics data types can explain the changes which lead to disease, i.e., the cause of disease and thus the treatment targets. Further this data can be used to molecular studies.

28.2 Strength of Individual Omics

28.2.1 Genomics

The word “genome” came from association of two words “gene” and “chromosome.” The complete set of hereditary information which is needed for function and development of a living organism is described by the word genome. The word genomics deals with the study of gene, i.e., a combination study of recombinant DNA, DNA sequencing technologies and bioinformatic analysis of sequences, assembly, and structural and functional analysis of genomes. The difference between genomics and “classical genetics” is that the former deals with full hereditary material, i.e., DNA of an organism, whereas the latter is the study of one gene or one gene product at a time. The entire genome of a free-living organism named

Haemophilus influenzae was sequenced in the 1990s which is considered a significant contribution to the field of genomics [6].

Sequencing of individual genomes for study of genomics has produced an immense knowledge on the respective strains. The comparative genomics provides more details of various organism specially dynamics of bacterial genomes. Comparative genomics is also very useful in clinical microbiology. One of those is metagenomics which is the study of genetic material that recovered directly from environmental samples. This study provides the knowledge of taxonomical diversity in environment which is uncultivable as well as provides insight of gene pool of clinically important gene such as resistance gene. Another field of study is epigenomics which deals with functional analysis such as modification of gene like methylation and histone modifications—acetylation, phosphorylation, methylation, and ubiquitination. Pharmacogenomics is a field of study which is a bridge between genomics and pharmacology. This field examines the response of drug in individual and its inheritance, thus optimizing drug therapy. They help in evaluation of rigorous systemic toxicity and unpredictable efficacies of therapeutics in the field of oncology. These technologies are valuable to identify novel targets for the treatment for various complex diseases such as cancer, cardiovascular disease, and obesity. Thus in the future, these technologies used in systems biology promise to develop new approaches to predict and prevent diseases [7]. By taking the advantages of such technologies, the research of obstetrics and gynecology is trying to solve the problem of infertility.

28.2.2 Application of Genomics

28.2.2.1 Design of Polymerase Chain Reaction (PCR) Assays

Polymerase chain reaction (PCR) was developed by Mullis in the 1980s [8]. Since the discovery, it is a widely used technique to detect microorganism in clinical specimens. In the early genomics era, genomic sequences empowered the selection of PCR primers of targeted DNA fragments which are specific to genus, species, subspecies, or strain depending on the objective [9, 10]. Though PCR is a basic technique of molecular biology or genomics, with time many advanced PCR techniques also came into use. With the advancement of genome sequencing and deep knowledge of genome sequences, it helped to design multiplex PCR. Multiplex PCR assays are useful to simultaneous detection of various microorganisms. As an example, a multiplex real-time PCR assay was designed from knowledge of genes to identify members of the *Mycobacterium tuberculosis* complex (MTC) with simultaneous differentiation between *M. tuberculosis* and *Mycobacterium canettii*. Other than real-time PCR, overlap extension PCR to insert mutations at specific points or to join the spliced DNA fragments into a larger one also exist. Droplet PCR is used to enhance the ability of PCR where each oil droplet acts as individual PCR reaction than each tube which increases the capacity of amplification of gene drastically. PCR is applied usually with known sequence. However, it can also be used in cases where we want to explore the unknown flanking regions. Restriction enzyme site is

incorporated in the unknown DNA sequence, and then using that sequence as primer unknown gene is amplified.

28.2.2.2 Genotyping

Genotyping is a traditional typing method to discriminate between bacterial species. Biotyping and serotyping are a process of genotyping which can discriminate within species. Further the molecular typing can be sequence based or non-sequence based. For example Non-sequence-based genotyping methods are based on the size of the DNA separated in the gel electrophoresis such as pulsed field gel electrophoresis (PFGE), multiple-locus variable-number tandem-repeat analysis (MLVA), PCR-restriction fragment length polymorphism (PCR-RFLP), single-nucleotide polymorphisms (SNPs), and microarrays. The sequence-based genotyping methods are multispacer sequence typing (MST), multilocus sequence typing (MLST), and whole genome sequence typing (Fig. 28.1).

28.2.2.3 Reverse Vaccinology

Microbial genomics has also made evolution in vaccinology. Microbial genomes encode the complete property of a strain including its antigenic property. Thus information of microbial genome in turn helps to identify putative antigenic proteins which are surface exposed and well conserved between strains, and such information

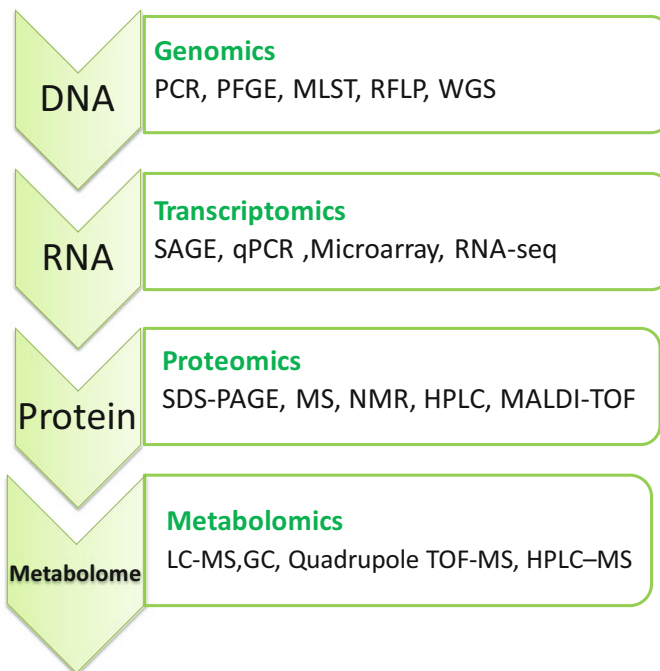


Fig. 28.1 Diverse fields of omics technology

can be used to acquire first information for vaccine development initially. The process is known as “reverse vaccinology” [11] and it was first proposed in 2000. The starting idea was based on the identification of novel meningococcal vaccine candidates by analyzing the genome sequence of *Neisseria meningitidis* serogroup B. There are also example of studies where reverse vaccinology is used such as study of organisms like pathogenic *E. coli* [12] and also in *Streptococcus agalactiae* [13], in addition with other organisms *Bacillus anthracis*, *Chlamydia pneumoniae*, *Leptospira interrogans*, *M. tuberculosis*, *P. gingivalis*, *Rickettsia prowazekii*, *S. pneumoniae*, and *Streptococcus pyogenes*.

28.2.2.4 Genomics-Based Design of Culture Media

It is a well-known fact that 99% of microorganisms found in the natural environments are not cultivable using laboratory techniques available till date. With the development of genomics, there are culture-independent techniques which are able to enlighten our knowledge of the diversity of microorganisms. Still cultivation is a crucial process before stepping into other processes like diagnosis, characterization of the pathogens, and antibiotic susceptibility profiling. Besides all these, the significant interest in the basic microbiological science is to culturing the unculturable. But it remains a major challenge of microbiology today as micro-organism can be fastidious and either partial or total lack of a metabolic pathway is among the reasons behind not able to culture in vivo. Genomic sequences have given access to the complete knowledge of metabolic potential of a strain which may help identify what is missing in the culture media which is important as per the genome sequence of the species. Thus, according to the metabolic need of that organism, culture media can be modified and optimized. As an example, the genome of *Tropheryma whipplei*, which is causative agent of Whipple’s disease [14, 15], lacks genes that are involved in biosynthesis pathways of nine essential amino acid. Thus the culture medium is specifically designed to incorporate all nine amino acids which in turn enable the growth of bacterium [16, 17]. Same kind of approach is also used for *C. burnetii* [18].

28.2.2.5 Detection of Virulence Factors

In order to understand the cause and develop specific treatment of a disease, it is primary step to identify and characterize the virulence factors of a pathogen and thus the pathogenesis of the diseases. There are various strategies that have been developed with the help of genomic sequencing technology in order to detect gene encoded for virulence factor, including: (i) comparing of genomic information of strains or species which exhibit various degree of pathogenesis; (ii) studying the genomic islands which give insight of those gene that are likely acquired due to lateral gene transfer; or (iii) identifying the virulence marker by searching the genomes of interest within known database. For example an approach first time used to compare between *Y. pestis*, causing plague with *Yersinia pseudotuberculosis*, lesser virulent but closely related species [19]. Another example is the identification of virulence factor by the study of comparing between *E. coli* O157:H7, a pathogenic strain, and *E. coli* K-12, a non-pathogenic counterpart used in laboratory

[20, 21]. Another study revealed the virulence of *Staphylococcus epidermidis* causing community-acquired endocarditis in comparison with commensal strains [22]. Uropathogenic *E. coli* strain genome is a second example where genomic island is found which encompassed 13% of the complete genome. [23]. In another study of *E. coli* strain PA45 829 genes identified had matches when compared with virulence factor database [24].

28.2.2.6 Detection of Antibiotic Resistance

Antimicrobial resistance is a growing public health problem. Besides conventional detection, antibiotic resistance can be predicted by analyzing the genomic content and adapt a specific treatment. Same strategy can be implied to identify resistance markers as for virulence genes [25]. A study reported from France describes that comparative genomic analysis identified a large “resistance island” containing 45 antimicrobial resistance-encoding genes by comparing genome of a multidrug-resistant and a wild-type *Acinetobacter baumannii* strains [26], whereas in Denmark, searching database approach was applied to identify 14 antibiotic resistance-associated genes in an epidemic strain using ResFinder and ARDB databases [27]. With the evolving time, whole genome sequencing (WGS) utilize for surveillance of antimicrobial resistance routinely compared with phenotypic procedures, Zankari et al., by using sequencing technology like Illumina [28].

28.2.2.7 Role of WGS in Epidemiology

In clinical microbiology laboratory, it is routine work to detect antibiotic susceptibility and virulence gene. But on the other hand, those labs frequently deal with outbreak detection and epidemiological investigations of microbes by WGS [29, 30]. One study in 2010 demonstrated the discriminatory power of WGS with the study of 63 strains of methicillin-resistant *S. aureus* (MRSA) collected from various countries where by analyzing genome data, they have described intercontinental transmissions and in addition it demonstrated transmission within hospital environment [30]. There are several examples of using WGS to investigate epidemiology, including study from Haiti in 2010 regarding cholera outbreak ([29]) and *E. coli* O104:H4 outbreak in Germany and France in 2011.

28.2.3 Transcriptomics

The word “transcriptome” first coined in the early 1990s [31] represents all transcripts that are present in a cell that includes mRNA, miRNA, noncoding RNAs, and small RNAs. As the name suggests, transcriptomics is the study of transcripts such as quantity of RNA, structure of transcribed RNA, and quantification of differential expression levels of transcripts at various stages of development of an organism under various physiological conditions. This kind of study refers to diversity of transcript, noncoding RNAs, and also the arrangement of coding portion of a transcript. The importance of transcriptome is to interpret essential functional elements of the genome and to study constituent molecule of a cell and tissues to

understand the disease development. Various technologies have been developed for the study of transcriptomics which include hybridization-based approach or sequence-based approaches which are described below. The early techniques of transcriptomic analysis were based on Sanger sequencing called EST, i.e., expressed sequence tags, and then another technique called SAGE, i.e., serial analysis of gene expression, came into existence. EST and SAGE were laborious process and able to determine only a small set of transcripts that are also in random fashion, thus being capable to yield only half information of transcriptome. Then advanced techniques like microarray and RNA-Seq were developed.

28.2.4 Application for Transcriptomics

28.2.4.1 Expression Sequence Tag (EST) and Serial/Cap Analysis of Gene Expression (SAGE/CAGE)

EST is a technique based on Sanger sequencing and generates short oligonucleotide sequence. In this technique RNA is first transcribed to cDNA using reverse transcriptase enzyme followed by sequencing of cDNA generated. EST is comparative simple technique and can be generated from any mixture of samples because the techniques do not demand prior knowledge of the origin of sample. ESTs are unique sequences capable of pointing the expressed genes of the mapped cDNA clone. This knowledge can be utilized to identify unknown gene but is unable to quantify the expression of genes. Another drawback of EST was low-throughput and costly method as it can sequence a single cDNA copy at a time. But with the evolving time, EST is not in use. Serial analysis of gene expression (SAGE) was an advanced version of EST. The technology SAGE was invented by Dr. Victor Velculescu from Johns Hopkins in 1995. Unlike EST in this technique quantitation of gene expression can be done and also identifies novel gene expression in a cell population. But the drawback was although gene expression is possible, actual gene expression cannot be measured. Besides this gene quantification can be biased due to linker dimer molecular contamination which leads to error in sequencing.

28.2.4.2 Microarray

The discovery of microarray technique in 1990 has brought revolution in the field of transcriptomics. This technique initiates the approach to predict biological functions with the help of the study of functional related genes and how it is globally expressed in combination with pathway analysis. It yields a large amount of data which decipher huge information related to biological activity of a cell. Initially the technique was developed to quantify multiple gene expressions simultaneously in a given time period. But as the time passes with scientific inventiveness, the technique has developed from two-dimensional to three-dimensional microarray and further followed by suspension bead arrays. These are now used largely in clinical settings. Microarray has adapted for global transcript analysis which is basically originated from mapping of genomic DNA. Till date there are several

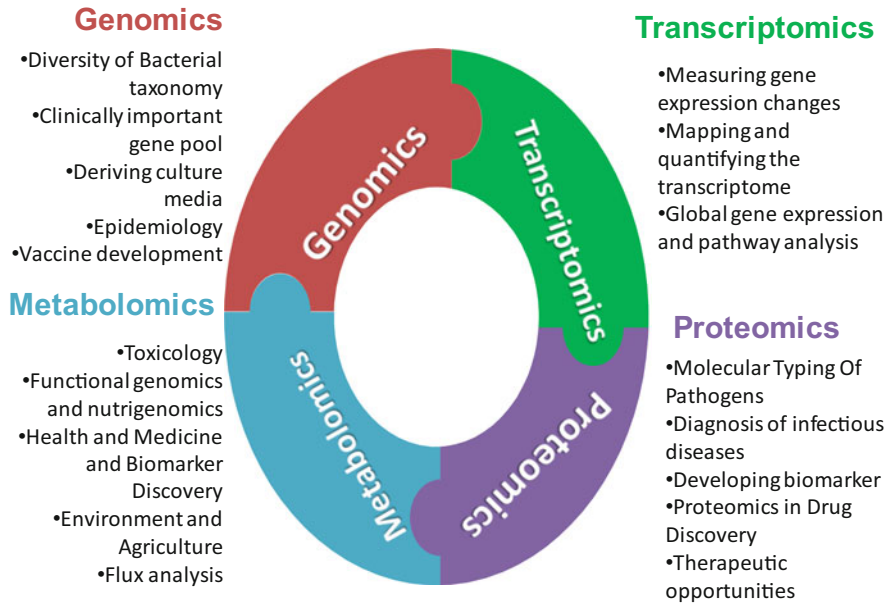


Fig. 28.2 Applications of omics technologies

other types of microarray which are under development that have potential to bring improvement in medical treatment as well as research (Fig. 28.2).

The basic principle behind microarray is hybridization which occurs between complementary strands of DNA. Among the complementary strands of DNA, one strand is called probe which is arrayed on microchip, and the other is the target transcript which is fluorescently labeled. Depending on the fluorescent intensity of each probe, abundance of transcript of a gene or RNA was determined, whereas the target was identified by the position on the chip. As a result microarray is capable to produce both quantitative and qualitative data. The former generates information of gene expression, whereas the latter is useful in the field of diagnosis. Microarray chip consists of two phases: the first is the immobilized phase, in which probes are prepared from cDNA or genome sequence or oligonucleotides, and the second is mobile phase which contains labeled cDNA.

Microarray is a widely used technique in research for measuring gene expression changes and elucidating the relationship between genotypes and phenotypes. It is able to analyze large mammalian transcriptome rapidly such as the gene expression alteration in a cancer cell. Data generated by microarray are used to analyze thousands of genes from multiple samples which are useful in drug development and clinical research. It is also used in clinical diagnostics. In some example molecular signatures specific to differentiate between acute myeloid leukemia and acute lymphocytic leukemia are identified. Another study by a group characterized molecular signature which can distinguish between following disease, neoplastic prostate, localized cancer, and metastatic prostate cancer from nonneoplastic and

neoplastic prostate tissues and healthy individuals. Microarray is also used in drug resistance field to identify the cellular pathways which are responsible for resistance, and thus this information can be used in overcoming the drug resistance. The drawback of the microarray method is that it cannot identify novel transcripts.

28.2.4.3 RNA-Seq

RNA sequencing is a novel and advanced method in the field of transcriptomics and applied for mapping and quantifying the transcriptome utilizing deep sequencing approach. This is a cost-effective high-throughput analysis method of transcriptome. There are several advantages over microarray, but the major advantage is the discovery of novel RNA species, i.e., the technique is not limited to detect only those transcripts corresponding to existing genomic sequence. In contrast to microarray, RNA-Seq is capable to offer higher sensitivity, and dynamic range of expression is captured where the former exhibits saturation. RNA splice events can be detectable by RNA-Seq which is not possible through microarray [32]. RNA-Seq is capable to reveal precise boundary location of transcripts with a single-base resolution and also reveal connectivity between exons which make it useful for the study of complex transcriptomes [33].

28.2.5 Proteomics

Protein is a vital molecule which is involved in the cellular function. DNA contains the blueprint of a cell, and proteins serve as ultimate building blocks. Proteome can be defined as “the complete protein complement expressed by a cell or unicellular organism at any one time” and the name “proteome” was coined by Marc Wilkins in 1994. The proteome can be described as dynamic entity which keeps changing, qualitatively and quantitatively, depending on the metabolism of cell in response to both intracellular and extracellular environment or mutations in DNA. So, following “genomics” and “transcriptomics,” the next major omics approach is to study proteome which is known as “proteomics.” The term “proteomics” was coined in 1997. This field of study includes exploration of the proteomes including the study of protein structure, composition, and function. Proteomics is complex than the other because an organism’s genome can be considered as more or less constant, but the proteome of organism is highly dynamic as it differs with conditions and time. In recent years with the development of technologies, proteomics is being used in many different areas of research in biological sciences including microbiology.

The primary interest of a medical microbiologist is in disease-causing potential of pathogenic organisms and the interaction with their hosts to achieve the goal of developing treatment of preventive measure of infectious disease. Genomics alone cannot provide answer to the entire question and thus transcriptomics evolves to determine its protein content. But it is also found inaccurate as it does not have correlation with protein expression [34, 35]. The mRNA is not directly correlated with the protein content because of posttranscriptional modifications, alternative splicing, mRNA editing, and proteolysis.

28.2.6 Applications of Proteomics

28.2.6.1 Molecular Typing of Pathogens

For the taxonomic and epidemiological investigations in medical microbiology, it is essential to discriminate between isolated species. Thus proteomics is used in the molecular epidemiology of infectious agents by detecting specific markers which can be further used to track distribution and movement of infectious agent within host population. There are varieties of techniques to identify suitable markers such as one-dimensional gel electrophoretic (1DGE) which can be extended by two-dimensional gel electrophoresis (2DGE). The non-fractionated cellular proteins resolved by 2DGE in a comparative study that is used to differentiate clinical isolates of bacterial groups such as *Campylobacter* sp. and *Neisseria* sp. are well studied.

With the ever-evolving technologies, the improved methods which are being recently introduced are used to differentiate between bacterial isolates by characterizing specific molecules with high-resolution mass spectrometry. One of these techniques is MALDI-TOF MS where bacterial isolates are compared as their surface protein component. As an example, one of the studies reported discrimination between methicillin-resistant *Staphylococcus aureus* and its sensitive counterparts with differed peaks with masses that range from 500 to 3500 Da. Another aspect of proteomics is the comprehensive study of protein, for example, the comparison between *M. tuberculosis* (two virulent strains) and *M. bovis* (two avirulent vaccine strains) using 2DGE and peptide mass mapping.

28.2.6.2 Diagnosis of Infectious Diseases

Tuberculosis is a disease which affects millions of people in the world where infection with drug-resistant organism is difficult to treat. A serum screening test is able to detect preclinical infection which in turn helps to give early treatment by potentially reducing transmission and thus is in widespread application. A study reported that such serum analysis of antigen has a potential to diagnose disease with sensitivity of 60%–74% and specificity of 96%–97% [36]. In another study of comparison between patients with SARS and healthy individuals, proteomic analysis revealed potential of truncated protein antitrypsin as a biomarker of the disease [37].

28.2.6.3 Developing Biomarker

As mentioned previously, the aim of medical microbiology is to understand pathogenicity and host-pathogen interaction. There are a number of protein determinants coded by pathogen, and they act in development of pathogenicity and help microorganism to colonize within host, followed by interaction, and finally surviving within the host. Proteomics play an important role to identify such determinants. As protein expression alters during disease condition in biological pathways, it can be monitored in tissue, blood, urine, or other biological samples which can provide indicators for the disease [38]. Expression of proteomics provides biomarker detection through comparison of protein expression profile between normal samples and disease-affected ones. The simplest approach used in biomarker discovery is

2D-PAGE in which protein profiles are compared between normal and disease samples such as tumor tissues and body fluids [39]. Approaches like MALDI-TOF and surface-enhanced laser desorption/ionization (SELDI)/protein chip techniques are used to identify biomarkers for different diseases [38]. According to information which they provide, disease-specific biomarkers can be divided into diagnostic, prognostic, and treatment-predictive biomarkers [40]. A diagnostic biomarker is used for early detection or presence of the disease. A prognostic biomarker usually is used to predict the recurrence and aggression of disease and a patient response to treatment by a given drug. This classification also is important in drug design applications [41]. It is estimated that only 2% of human diseases appear due to single gene damage. Other factors like epigenetic and environmental factors involved in the development and outcome of the diseases account for the remaining 98% [42]. In this regard, proteomics can be helpful to identification of proteins that can potentially serve as disease-associated biomarkers which are involved in disease progression. After identifying biomarkers by mass spectrometry-based approach, biomarkers need to be processed using bioinformatics analyses and also need to be reproduced in different populations [43]. The proteomics biomarker discovery is advanced in a variety of diseases such as cancer [44], cardiovascular diseases, acquired immune deficiency syndrome (AIDS), renal diseases, and diabetes [45].

28.2.6.4 Proteomics in Drug Discovery

Proteomics study based on MS has expanded over time, and its role in almost all diverse research fields of science has developed. As the drug discovery is an inherently complex process and values high cost, new emerging technologies such as proteomics can facilitate and accelerate discovery processes. It is estimated that finding each new drug candidate costs \$70 million [46]. Drug discovery has many stages which have been presented, and indeed it is a multidisciplinary field using genomics, proteomics, metabolomics, bioinformatics, and system biology. Proteomics studies also are useful for drug action, toxicity, resistance, and its efficacy under examination. By global searching of the proteome in a given sample such as tissue or cells treated with a drug, proteomics provides insights about disease-related molecular and cellular mechanisms that facilitate drug target discovery. Evaluated protein targets of *Lavandula angustifolia* on the treatment of rat Alzheimer's disease using 2D-PAGE-MALDI-TOF/TOF. Several techniques are used for target and lead identification [47] such as isotope-coded affinity tags (ICAT) [48], isobaric tags for relative and absolute quantification (iTRAQ) [49], and protein arrays [50]. The various MS-based platforms, including clinical, functional, and chemical proteomics, are also involved in the modern drug discovery process.

28.2.6.5 Therapeutic Opportunities

It is a well-known fact that cancer is a growing disease of modern life; thus understanding the cause of the disease helps to improve the existing therapy as well as to find out new treatment strategies. One study of proteomics recently reported that ten proteins in different cells with various structures and quantities are resistant to vinca alkaloids. Vinca alkaloids are sensitive cytoskeletal proteins

(e.g., tubulin and actin) or an entity that is able to bind to cytoskeletal proteins like heat shock protein [51]. With the development of omics technology, genome and protein sequence data of many microorganisms are available. Those data are useful to provide knowledge for understanding their resistance to drugs which in turn help to develop tools to identify drug resistance pattern and novel molecules which can be used to treat drug-resistant disease. One such example is azole resistance in *Candida albicans* which has a differential expression of protein involved in the ergosterol biosynthesis pathway which can be used as drug target [52, 53].

28.2.7 Metabolomics

The central dogma of passing information in a cell starts from gene to transcript then to protein and metabolites. As the name suggests, metabolomics is a high-throughput study of metabolites which can be defined as integral part of metabolism process. Metabolomics deals with metabolites with a molecular weight of < 1500 Da, such as carbohydrate, fatty acids, and hormones, which can vary during metabolism in response to internal or external factors with various time [54, 55]. Metabolome of organism is complex and dynamic as proteome because with time metabolites are absorbed and degraded as it is continuously synthesized. So the metabolomics studies are an attempt to provide a picture of physiological state of organism with a specific time. Targeted and untargeted are the two approaches which can be used to study metabolites. In untargeted approach, we used to study without any sample bias and determine the number of different metabolites, whereas in the case of targeted approach, defined sets of metabolites are used to measure to address a problem. However, in both the approaches, steps are the same.

After designing the study considering the factors like sample size, randomization, etc., the next step is sample preparation which includes sample collection, storage, and preparation. The third step is the technical part where techniques like mass spectroscopy or NMR are utilized to quantify the metabolites. Before analysis of data, it is crucial that quality control analysis of the data background correction is performed in order to avoid noise in the data to extract biological inferences. Finally, analysis of data is taken place with the application of statistical analysis and clustering of data. Metabolomics is an interdisciplinary field; thus it merged the data of various omics field such as genomics and proteomics to get a complete snapshot of metabolome study. The focus of the metabolomics is to study the enzymes, substrate, and product, thus forming a complex biological reaction. But in a biological system, this study is influenced by intrinsic factor such as genes and outer factor such as environmental conditions.

28.2.8 Applications of Metabolomics

There is a diverse range of application of metabolomics, and they can be considered among important tools for the study of living organisms in addition to diseases and

disease mechanisms. Moreover as growing technology, metabolomics are becoming more diverse including applications in identification of disease, drug discovery and drug development, nutrigenomics, and research in agriculture.

28.2.8.1 Toxicology

This is a well-studied area where metabolomics is widely used. To determine the toxicity level in body fluid such as urine and blood, this technique is used. Metabolomics also can be used to analyze liver- or kidney-associated disease condition.

28.2.8.2 Functional Genomics and Nutrigenomics

Metabolomics have great utility in study phenotype which is a result of change in gene and also are considered in the field of functional genomics. Elaborately knowledge of metabolome can suggest any change in gene. Due to deletion or insertion, gene may change, and a significant change can be seen in the production of metabolome. It is also able to predict function of unknown gene by comparing the metabolome profile of organism. These comparisons are currently undertaken in model organisms like *Saccharomyces cerevisiae* and *Arabidopsis* and could lead to advanced study in the future [31]. In a similar way, it can be used to determine nutrition utilizing profile of an organism. Another field of study where metabolomics can be applied is nutrigenomics which deals with the metabolic fingerprint of the organism, and by the help of metabolomics study, a whole snapshot can be determined depending on factors as well as individual system.

28.2.8.3 Health and Medicine and Biomarker Discovery

With the help of metabolomics, we can determine the metabolic changes in disease condition. Thus it has also a role in understanding the disease mechanisms such as cancer. Another application of this field is to study pathophysiological states of disease. Thus metabolomics can identify biomarker to categorize the progression of several types of diseases such as cancer. Biomarker identification is considered to be an important area of research in disease diagnostic. Using metabolomics, metabolite biomarker can be identified by comparing the metabolome profile of the disease group and the healthy group. Biological samples such as bile, urine, or seminal fluids are rich in information of metabolites. To identify biomarkers, these samples can be processed for information by metabolic profiling or fingerprinting, techniques used in metabolomics.

28.2.8.4 Environment and Agriculture

Metabolomics can also be applied to characterize interaction between organism and environment. In the study of change in metabolomic profile, we can identify the environment and organism interaction, and it can reveal information how environment affects organism's health. This can likewise be applied to a more extensive population to give information to different fields of exploration, for example, ecology. Metabolomics also can be used to improve the genetically modified crops. It can get a view of plant development at different time frames, thus being

capable of identifying loss caused by GM crops upon consumption. In the course of identifying the functions of primary and secondary metabolites, it is important to identify plant metabolite.

28.2.8.5 Flux Analysis

Metabolomics is a tool to study small molecules and metabolites at a given time point of a sample and evaluate them. These metabolites can be studied in an unbiased way with nontargeted approach with the help of techniques like mass spectroscopy and NMR. Mass spectroscopy can be used in combination with chromatography and analyze spectra with statistical techniques. But some metabolites are present in very low concentration in a sample, and thus there is a need of technology like flux balance analysis (FBA). This method is a limitation-based approach by enhancing the enduring condition and framework such as ATP generation. The steady-state assessment is finished by FBA by making use of the stoichiometric cross section for the structure.

28.3 Integrated Omics in Microbial System

28.3.1 Integrated Transcriptomics and Proteomics

With the difference in study question and design, there are various approaches to use integrated multiomics data [56]. Based on the central dogma of molecular genetics, it can be that there will be high correlation between mRNA expression levels and protein abundance as in central dogma protein is synthesized after mRNA synthesis. But it is not so and recent studies can fail to find significant correlation among transcriptomics and proteomics data [34] or be able to establish weak correlation [35, 57–59]. This discrepancy may rise due to several factors, such as changes of protein due to posttranslational modification, or self-modification, the differences of half-lives between mRNA and proteins, possible analysis required for protein binding, and also significant error in experiments [57, 60, 61]. So this observation also emphasizes on the need of integrated omics study, i.e., transcriptomics combined with proteomics, as an individual may fail to portray complete data [61]. The most common approaches among integrated omics are integrated approach of transcriptomics and proteomics. Integrated transcriptomics and proteomics data complement each other to detect the biasness of individual technologies and are able to provide better view of the metabolic changes. Another importance of such study is cross-validation of data. Conclusion can only be drawn confidently when high level of correlation or consensus is found between RNA and protein [62, 63]. Integrated transcriptomic and proteomic data can reveal a novel biological insight that is impossible for a single omics technology. For example, a study suggested possible posttranscriptional regulation in *Bacillus subtilis* due to chill adaptation [64] and physiological changes of *Halobacterium salinarum* with the change of oxygen level [65].

28.3.2 Integrated Transcriptomics and Metabolomics

Another important integrated omics study is integrated transcriptomics and metabolomics analysis. This integrated tool is powerful to study transcripts and the functional metabolites and relation between them. This kind of relation first showed that parallel analysis can be possible of transcript and metabolic profiles, and functionally importance of metabolites can be identified by the correlation of fungi and plant [66]. Another example of such integrated omics is transcriptomics, metabolomics, and fluxomics that were integrated and analyzed for lysine-producing *Corynebacterium glutamicum* ATCC 13287 of batch culture [67].

28.3.3 Tools Available for Integration of Multiomics Data

It is a computationally complex process to analyze thousands of measurements in each omics experiment where extraction of meaningful correlations and true interactions is needed. Biological frameworks frequently yield non-direct connections and joint impacts of numerous components, making it hard to recognize genuine natural signs from arbitrary noise. There are many sources to generate noise such as natural biological frameworks, inconsequential scientific stages, and different information explicit investigation work processes. Abundance of gene, protein, and metabolites largely varies among cell type, tissue type, and organ type. Thus organization can pose challenges for extraction of such data at biological level. With the increasing time, there are a number of studies incorporated in the field of omics with adverse array; some name of such newer omics approaches are fluxomics, ionomics, microbiomics, and glycomics. These are involved in the study of biomedical fields to identify and predict health status. Statistical and even machine learning tools are often required for integrative study and for multiomics view. To analyze integrated omics data, machine learning approaches are useful. It can be used for various approaches like clustering, identifying association between clinical measures, and also predicting disease [9]. Simplistic, descriptive, and exploratory approaches are taken for the analysis of data. Multivariate analysis tools like principal component analysis (PCA) decrease information dimensionality, where canonical correlation analysis (CCA) is used to explore the general relationship between two arrangements of variables. Other analysis types such as multivariate partial least square regression analysis and multiple factor analysis are also used in integrative omics study.

A recent study by showed a few accessible devices and bundles investigation of coordinated omics datasets utilizing pathway enhancement, organic organization, or observational connection examination. Yet, among those the greater part of the instruments require standard R-statistical programming or Python or Galaxy, usage of which has been characterized as “troublesome” by the creators however recommending the requirement for more easy to understand apparatus remains. As of now there are some web based devices are accessible for reconciliation of omics requires no computational experience just as more flexible tools for those with

computational experience. Such tools include Paintomics, 3Omics, and Galaxy (P, M) [68], though using tools blindly leads to adverse effect on data analysis. Other tools for more advanced users with expertise in programming such as IntegrOmics, SteinerNet, Omics Integrator, and MixOmics are available [68].

28.4 Current Challenges and Future Directions

28.4.1 Study Design

To gain knowledge of disease by omics technology mostly depends on comparative analysis. Most used methods use to compare the data between healthy and disease individuals, and the analyzed data is correlated with disease. However, there are many factors such as population structure, cell type composition, batch effects which may act as confounding factors in order to ascertain the result, and other unknown factors which lead to complex phenotypes of both “healthy” and “disease.” Mostly we do not know about the confounding factors assuming sources of variation may be due to sex, BMI, age, and diet in metabolic disease. This problem can be overcome if sample amount is large. Thus, there is a need of collecting accurate large dataset to capture the source of variance in background population which is one of the crucial aspects for the success of omics studies.

In the case of transcriptomics, there are many steps such as cDNA synthesis from extracted RNA and library preparation. After adapter ligation small RNAs (microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), short interfering RNAs (siRNAs)) and numerous others can be straightly sequenced. But problems arise in the case of larger RNA molecules. Before sequencing for large RNAs, it must be fragmented into short length pieces (200–500 bp) that it could be suitable for sequencing with most deep sequencing technologies. To fragment RNA the most common fragmentation methods used is by RNA hydrolysis or nebulization, and in the case of cDNA, fragmentation method used is by DNase I treatment or sonication. But these methods lead to biased fragmentation. But there is biasness, for example, RNA fragmentation has little bias than cDNA fragmentation. cDNA is normally firmly one-sided toward the ID of groupings from the closures of records [32].

28.4.2 Analytical Challenges

The major advantage of large omics datasets is availability of data for reanalysis with multiple approaches again. By using such enduring availability of data, it is important for omics fields to develop statistical methods that can extract more information from the data that already exist. While every omics field deals with its own challenges, one common challenge is the analysis of data. For a complete knowledge, one cannot depend on one kind of omics, and thus multiomics data analysis come in action and analysis of which is difficult due to establishing its correlation.

All high-throughput sequencing faces the challenges such as improvement of productive techniques for proper storage, retrieval, and processing such huge amount of data. The challenges of data analysis need to be addressed in order to reduce error of image analysis and base calling, as well as it is important to quality control check the data. In the case of transcriptomics, poly A tail can be identified by searching the repeated presence of As or Ts at the end of some reads. Exon-exon junctions can be identified with specific sequence context of GT-AG dinucleotides flanking at splice sites. Further low expression of intronic sequences are removed during splicing in *S. cerevisiae* [69].

28.4.3 Coverage Versus Cost

Though with time cost of high-throughput sequencing has lowered significantly, it varies with the coverage. To analyze the data and overcome the issues described above, higher coverage is needed but greater coverage requires more sequencing depth.

28.4.4 Challenges of Transcriptomics

RNA-Seq has the potential to generate single-base resolution which can revise annotation of existing gene including gene boundaries and introns. It is also capable of identifying new gene transcripts. RNA-Seq with precipitous drop in signal can be used to map 5' and 3' boundaries within 10–50 bases. In addition it can precisely map 3' boundaries via looking for poly(A) labels or via looking through labels that length GT-AG grafting agreement locales of introns. As an example in *S. cerevisiae*, boundaries of 80% of all annotated genes were mapped by using these methods. Similarly, many boundaries are defined in *S. pombe* by RNA-Seq data in combination with tiling array.

28.4.5 Challenges of Proteomics and Metabolomics

Low abundance is one of the major problems associated with proteomics as the low-abundance proteins such as transcription factors, kinases, and regulatory proteins are difficult to detect. It is difficult to detect such low-expressing proteins and crude cell lysates without sophisticated purification methods. But unfortunately there are no protein or metabolite amplification techniques available in the case of genomes, i.e., PCR. So it is burdensome to study proteins and metabolics on a scale of genomics. In addition methods used in proteomics and metabolomics are not high throughputs, although MS is widely used for proteomics and metabolomics to identification and analysis which is still time-consuming and often with risk of sacrificing quality. Though MS is capable of providing higher-quality data, challenges arise in the interpretation of data which is time-taking and suggest that

better computational algorithms is needed with increased accuracy of data interpretation without any manual intervention. Moreover such biomolecules are difficult to quantify.

28.5 Future Prospects and Conclusion

Studies in combination with genomics, transcriptomics, and proteomics can give us an overview from what is there in the cell to the actual happening inside the cell. Though the concept looks attractive, it comes with numerous challenges in data analysis. To get the real view of the data, it is needed to overcome the challenges. There are also some problems which are associated with the systems of the approach as follows: (i) as different platforms have different way to behave, it is difficult to analyze dataset in integrated omics, making it difficult and almost impossible to analyze integrated datasets of transcriptome and proteome; (ii) it is also difficult to normalization of data across different platforms; (iii) different approaches produce data in different file formats, and thus it is a huge challenge to deal with and harmonizing these; and (iv) another challenge is to annotate DNA, RNA, and proteins as still there are some genes whose product is unknown. Currently, there is no single approach to process, analyze, and interpret all data formats from different omics and integrated analysis. Hence for the advancement, it needed a tool or algorithm for multimodal data combination strategies that can be reproducible, user-friendly, and effective frameworks with high throughput.

References

1. Smith MT, Vermeulen R, Li G, Zhang L, Lan Q, Hubbard AE et al (2005) Use of ‘Omic’ technologies to study humans exposed to benzene. *Chem Biol Interact* 153:123–127
2. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409(6822):928–934
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al (2001) The sequence of the human genome. *Science* 291(5507):1304–1351
4. Hasin Y, Seldin M, Lusic A (2017) Multi-omics approaches to disease. *Genome Biol* 18(1):1–15
5. Poisot T, Péquign B, Gravel D (2013) High-throughput sequencing: a roadmap toward community ecology. *Ecol Evol* 3(4):1125–1139
6. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269(5223):496–512
7. Sagner M, McNeil A, Puska P, Auffray C, Price ND, Hood L et al (2017) The P 4 health spectrum—a predictive, preventive, personalized and participatory continuum for promoting healthspan. *Prog Cardiovasc Dis* 59(5):506–521
8. Mullis KB, Faloona FA (1987) Specific synthesis of DNA in vitro via a polymerase catalysed chain reaction, in “Methods Enzymol” vol 155 R. Wu. ed.
9. Li W, Raoult D, Fournier PE (2009) Bacterial strain typing in the genomic era. *FEMS Microbiol Rev* 33(5):892–916

10. Yang JY, Brooks S, Meyer J a, Blakesley RR, Zelazny AM, Segre JA, Snitkin ES (2013) Pan-PCR, a computational method for designing bacterium-typing assays based on whole-genome sequence data. *J Clin Microbiol* 51:752–758
11. Mora M, Veggi D, Santini L, Pizza M, Rappuoli R (2003) Reverse vaccinology *Drug Discov Today* 8:459–464
12. Moriel DG, Bertoldi I, Spagnuolo A, Marchi S, Rosini R, Nesta B, Pastorello I, Corea VAM, Torricelli G, Cartocci E, Savino S, Scarselli M, Dobrindt U, Hacker J, Tettelin H, Tallon LJ, Sullivan S, Wieler LH, Ewers C, Pickard D, Dougan G, Fontana MR, Rappuoli R, Pizza M, Serino L (2010) Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A* 107:9072–9077
13. Maione D, Margarit I, Rinaudo CD, Massignani V, Mora M, Scarselli M, Tettelin H, Brettoni C, Iacobini ET, Rosini R, D'Agostino N, Miorin L, Buccato S, Mariani M, Galli G, Nogarotto R, Nardi-Dei V, Nardi Dei V, Vegni F, Fraser C, Mancuso G, Teti G, Madoff LC, Paoletti LC, Rappuoli R, Kasper DL, Telford JL, Grandi G (2005) Identification of a universal group B streptococcus vaccine by multiple genome screen. *Science* 309:148–150
14. Bentley SD, Maiwald M, Murphy LD, Pallen MJ, Yeats CA, Dover LG, Norbertczak HT, Besra GS, Quail MA, Harris DE, von Herbay A, Goble A, Rutter S, Squares R, Squares S, Barrell BG, Parkhill J, Relman DA (2003) Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*. *Lancet* 361:637–644
15. Raoult D, Ogata H, Audic S, Robert C, Suhre K, Drancourt M, Claverie J-M (2003) *Tropheryma whipplei* twist: a human pathogenic actinobacteria with a reduced genome. *Genome Res* 13:1800–1809
16. Fenollar F, Rolain J-M, Alric L, Papo T, Chauveheid M-P, van de Beek D, Raoult D (2009) Resistance to trimethoprim/sulfamethoxazole and *Tropheryma whipplei*. *Int J Antimicrob Agents* 34:255–259
17. Renesto P, Crapoulet N, Ogata H, La Scola B, Vestris G, Claverie J-M, Raoult D (2003) Genome-based design of a cell-free culture medium for *Tropheryma whipplei*. *Lancet* 362:447–449
18. Omsland A (2012) Axenic growth of *Coxiella burnetii*. *Adv Exp Med Biol* 984:215–229
19. Chain PSG, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, Regala WM, Georgescu AM, Vergez LM, Land ML, Motin VL, Brubaker RR, Fowler J, Hinnebusch J, Marceau M, Medigue C, Simonet M, Chenal-Francois V, Souza B, Dacheux D, Elliott JM, Derbise A, Hauser LJ, Garcia E (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci U S A* 101:13826–13831
20. Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic environmental microorganisms. *Nat Rev Microbiol* 2:414–424
21. Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick H a, Pósfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch R a, Blattner FR (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409:529–533
22. Fournier PE, Gouriet F, Gimenez G, Robert C, Raoult D (2013) Deciphering genomic virulence traits of a *Staphylococcus epidermidis* strain causing native-valve endocarditis. *J Clin Microbiol* 51(5):1617–1621
23. Lloyd AL, Rasko DA, Mobley HLT (2007) Defining genomic islands and uropathogen specific genes in uropathogenic *Escherichia coli*. *J Bacteriol* 189:3532–3546
24. Segata N, Ballarini A, Jousson O (2013) Genome sequence of *Pseudomonas aeruginosa* PA45, a highly virulent strain isolated from a patient with bloodstream infection. *Genome Announc* 1: e00289–e00213
25. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW (2012) Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet* 13(9):601–612

26. Fournier PE, Vallenet D, Barbe V, Audic S, Ogata H, Poirel L, Richet H, Robert C, Mangenot S, Abergel C, Nordmann P, Weissenbach J, Raoult D, Claverie J-M (2006) Comparative genomics of multidrug resistance in *Acinetobacter baumannii*. *PLoS Genet* 2:e7
27. Tan SY, Chua SL, Liu Y, Hoiby N, Andersen LP, Givskov M, Song Z, Yang L (2013) Comparative genomic analysis of rapid evolution of an extreme-drug-resistant *Acinetobacter baumannii* clone. *Genome Biol Evol* 5:807–818
28. Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agersø Y, Lund O, Larsen MV, Aarestrup FM (2013) Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. *J Antimicrob Chemother* 68(4):771–777
29. Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamaichi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK (2011) The origin of the Haitian cholera outbreak strain. *N Engl J Med* 364:33–42
30. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD (2010) Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327:469–474
31. Arivaradarajan P, Misra G (2019) Omics approaches, technologies and applications: integrative approaches for understanding OMICS data. Springer
32. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10(1):57–63
33. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK et al (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* 5(7):613–619
34. Gygi SP, Rochon Y, Franz BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 19:1720–1730
35. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292:929–934
36. Bahk YY, Kim SA, Kim JS, Euh HJ, Bai GH, Cho SN, Kim YS (2004) Antigens secreted from *Mycobacterium tuberculosis*: identification by proteomics approach and test for diagnostic marker. *Proteomics* 4(11):3299–3307
37. Ren Y, He QY, Fan J, Jones B, Zhou Y, Xie Y et al (2004) The use of proteomics in the discovery of serum biomarkers from patients with severe acute respiratory syndrome. *Proteomics* 4(11):3477–3484
38. He QY, Chiu JF (2003) Proteomics in biomarker discovery and drug development. *J Cell Biochem* 89(5):868–886
39. Gam LH (2012) Breast cancer and protein biomarkers. *World journal of experimental medicine* 2(5):86
40. Srinivas PR, Verma M, Zhao Y, Srivastava S (2002) Proteomics for cancer biomarker discovery. *Clin Chem* 48(8):1160–1169
41. Hamdan MH (2007) Cancer biomarkers: analytical techniques for discovery, vol 25. Wiley
42. Palzkill T (2002) Kluwer. Proteomics. Kluwer Academic Publishers, New York
43. Khadir A, Tiss A (2013) Proteomics approaches towards early detection and diagnosis of cancer. *Journal of Carcinogenesis & Mutagenesis* S14:002
44. Safaei A, Rezaei-Tavirani M, Sobhi S, Akbari ME (2013) Breast cancer biomarker discovery: proteomics and genomics approaches. *Iran J Cancer Prev* 6:45–53
45. Safari-Alighiarloo N, Taghizadeh M, Tabatabaei SM, Shahsavari S, Namaki S, Khodakarim S, Rezaei-Tavirani M (2017) Identification of new key genes for type 1 diabetes through construction and analysis of protein-protein interaction networks based on blood and pancreatic islet transcriptomes. *J Diabetes* 9:764–777
46. Myers S, Baker A (2001) Drug discovery—an operating model for a new era. *Nat Biotechnol* 19(8):727–730

47. Burbaum J, Tobal GM (2002) Proteomics in drug discovery. *Curr Opin Chem Biol* 6(4):427–433
48. Tuñón J, Martín-Ventura JL, Blanco-Colio LM, Lorenzo Ó, López JA, Egado J (2010) Proteomic strategies in the search of new biomarkers in atherothrombosis. *J Am Coll Cardiol* 55(19):2009–2016
49. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S et al (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 3(12):1154–1169
50. Hall DA, Ptacek J, Snyder M (2007) Protein microarray technology. *Mech Ageing Dev* 128(1):161–167
51. Verrills NM, Walsh BJ, Cobon GS, Hains PG, Kavallaris M (2003) Proteome analysis of vinca alkaloid response and resistance in acute lymphoblastic leukemia reveals novel cytoskeletal alterations. *J Biol Chem* 278(46):45082–45093
52. Hooshdaran MZ, Barker KS, Hilliard GM, Kusch H, Morschhäuser J, Rogers PD (2004) Proteomic analysis of azole resistance in *Candida albicans* clinical isolates. *Antimicrob Agents Chemother* 48(7):2733–2735
53. Schmidt FR (2004) The challenge of multidrug resistance: actual strategies in the development of novel antibacterials. *Appl Microbiol Biotechnol* 63(4):335–343
54. Chen Z, Li Z, Li H, Jiang Y (2019) Metabolomics: a promising diagnostic and therapeutic implement for breast cancer. *Onco Targets and therapy* 12:6797
55. Nalbantoglu S (2019) Metabolomics: basic principles and strategies. In *Molecular Medicine*. Intech Open
56. Pálsson B, Zengler K (2010) The challenges of integrating multi-omic data sets. *Nat Chem Biol* 6:787–789
57. Greenbaum D, Jansen R, Gerstein M (2002) Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* 18:585–596
58. Nie L, Wu G, Culley DE, Scholten JC, Zhang W (2007) Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. *Crit Rev Biotechnol* 27:63–75
59. Washburn MP, Koller A, Oshiro G, Ulaszek G, Plouffe D, Decui C, Winzeler E, Yates JR III (2003) Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 100:3107–3112
60. Beyer A, Hollunder J, Nasheuer HP, Wilhelm T (2004) Posttranscriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. *Mol Cell Proteomics* 3:1083–1092
61. Park SJ, Lee SY, Cho J, Kim TY, Lee JW, Park JH, Han MJ (2005) Global physiological understanding and metabolic engineering of microorganisms based on omics studies. *Appl Microbiol Biotechnol* 68:567–579
62. Lee JH, Lee DE, Lee BU, Kim HS (2003) Global analyses of transcriptomes and proteomes of a parent strain and an L-threonine overproducing mutant strain. *J Bacteriol* 185:5442–5451
63. Nunez C, Esteve-Núñez A, Giometti C, Tollaksen S, Khare T, Lin W, Lovley DR, Methé BA (2006) DNA microarray and proteomic analyses of the RpoS regulon in *Geobacter sulfurreducens*. *J Bacteriol* 188:2792–2800
64. Budde I, Steil L, Scharf C, Völker U, Bremer E (2006) Adaptation of *Bacillus subtilis* to growth at low temperature: a combined transcriptomic and proteomic appraisal. *Microbiology* 152:831–853
65. Schmid AK, Reiss DJ, Kaur A, Pan M, King N, Van PT, Hohmann L, Martin DB, Baliga NS (2007) The anatomy of microbial cell state transitions in response to oxygen. *Genome Res* 17:1399–1413

66. Askenazi M, Driggers EM, Holtzman DA, Norman TC, Iverson S, Zimmer DP, Boers ME, Blomquist PR, Martinez EJ, authors o (2003) Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat Biotechnol* 21:150–156
67. Kromer JO, Sorgenfrei O, Klopprogge K, Heinzle E, Wittmann C (2004) In-depth profiling of lysine-producing *Corynebacterium glutamicum* by combined analysis of the transcriptome, metabolome, and fluxome. *J Bacteriol* 186:1769–1784
68. Misra BB, Langefeld C, Olivier M, Cox LA (2019) Integrated omics: tools, advances and future approaches. *J Mol Endocrinol* 62(1):R21–R45
69. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320(5881):1344–1349